

VU Research Portal

Trek alle registers open!

Bakker, B.F.M.

2009

document version

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Bakker, B. F. M. (2009). *Trek alle registers open!* Vrije Universiteit.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Prof. dr. Bart F.M. Bakker

Trek alle registers open!

Rede in verkorte vorm uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Methodologie van registers voor sociaalwetenschappelijk onderzoek bij de Faculteit der Sociale Wetenschappen van de Vrije Universiteit Amsterdam op 26 november 2009.

Mijnheer de rector, dames en heren,

Ik heet u van harte welkom bij deze openbare lezing die ik uitspreek ter gelegenheid van het aanvaarden van mijn leerstoel aan de Vrije Universiteit. Ik wil het met u hebben over het gebruik van administratieve gegevens, ook wel registers genoemd, voor de productie van statistieken en voor gebruik in sociaalwetenschappelijk onderzoek. Waarom is het de moeite waard om als universiteit daarvoor een leerstoel in te richten?

Daarvoor kunnen verschillende redenen worden aangevoerd. In de eerste plaats neemt het gebruik van registerinformatie toe. Als je kijkt naar de officiële statistieken van het Centraal Bureau voor de Statistiek, dan waren vijftien jaar geleden statistieken met een register als bron nog maar mondjesmaat aanwezig; tegenwoordig is meer dan de helft van de sociale en ruimtelijke statistieken gebaseerd op waarneming via registers. Het enquêteren voor het verzamelen van gegevens wordt meer en meer vervangen door verzamelen van gegevens via al bestaande administraties.

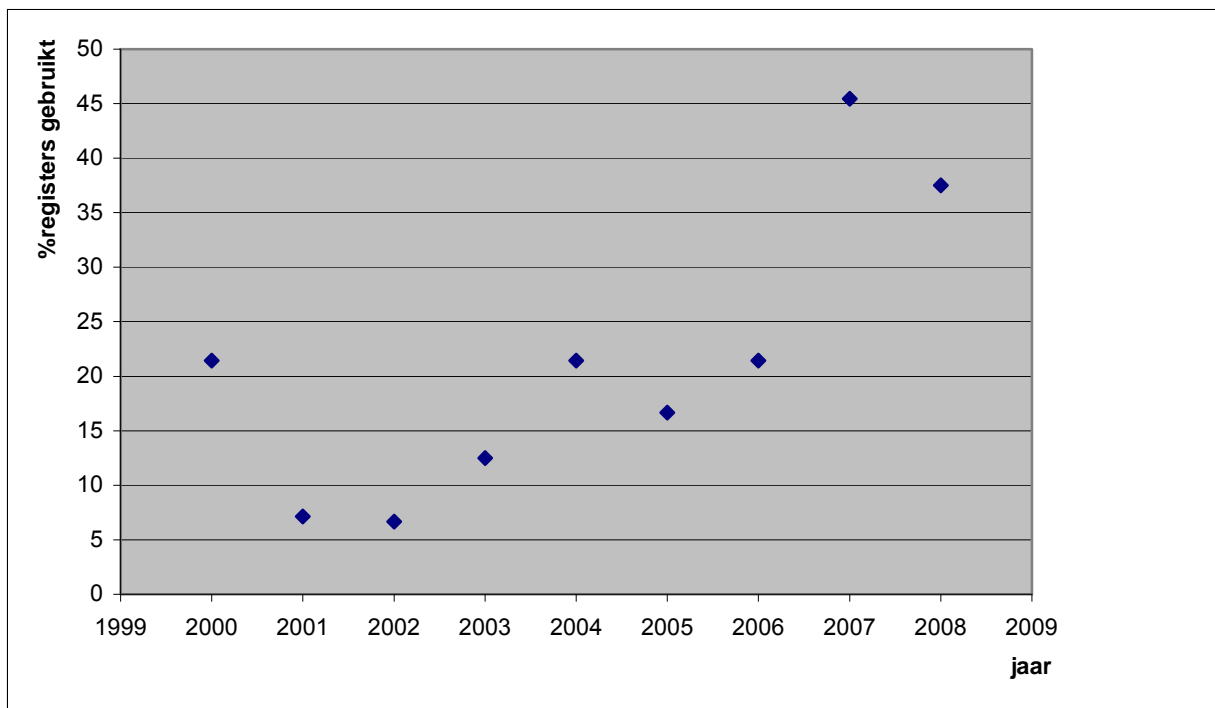
Overigens was ook voor de Tweede Wereldoorlog een groot deel van de statistieken gebaseerd op registers. Wat dat betreft herhaalt de geschiedenis zich. Registers zagen er wel anders uit: ze bestonden uit pakken papier in plaats van elektronische databestanden. In die tijd waren naast registers ook de Volkstellingen een belangrijke bron voor statistieken. De Volkstellingen werden eens in de tien jaar gehouden en leverden onder meer de belangrijkste demografische en sociaaleconomische informatie voor de samenleving. De ideeën om enquête-informatie (in die tijd Volkstellingen) door registerinformatie te vervangen dateren al van voor de Tweede Wereldoorlog. Methorst, die tot 1939 Directeur Generaal van het CBS was en daarna hoofd van de Volks- en Beroepstelling 1940, stelde voor om persoonsgegevens voor de Volkstelling te ontleen aan de nieuw op te zetten bevolkingsadministratie en alleen aanvullende gegevens te enquêteren. Door de Tweede Wereldoorlog is dit idee min of meer verloren gegaan (Van Maarseveen, 1999). Pas in deze eeuw heeft het CBS de Volkstelling van 2001 uitgevoerd door uitgebreid gebruik te maken van registers en al bestaande enquêtes. Daarmee heeft het de Nederlandse regering ongeveer 290 miljoen euro bespaard.

Ook voor het sociaalwetenschappelijk onderzoek wordt in toenemende mate gebruik gemaakt van registerinformatie. Je kunt dat in de tijdschriften voor sociaalwetenschappelijk onderzoek terugvinden. Zonder de pretentie

te hebben om compleet te zijn, wil ik dat illustreren aan de hand van het gebruik van registerinformatie in artikelen uit het sociaalwetenschappelijke tijdschrift *Mens en Maatschappij*. Als je teruggaat naar het begin van deze eeuw, dan wordt nauwelijks gebruik gemaakt van registers, terwijl in 2007 en 2008 al meer dan een derde van de artikelen gebaseerd is op registerinformatie. In 2000 zijn er drie artikelen waarin registers worden gebruikt. Een artikel over zelfdodingen waarin papieren telkaarten worden gebruikt, een artikel over etnisch gemengde huwelijken in Joegoslavië en een waarin registerdata van het CBS op buurniveau wordt gekoppeld aan een enquête. Vooral de eerste twee artikelen zijn atypische voorbeelden van registergebruik zoals je dat in latere artikelen tegenkomt. In latere artikelen worden elektronische registerdata gebruikt die in de meerderheid van de gevallen door het CBS ter beschikking zijn gesteld.

Figuur 1.

De ontwikkeling van het gebruik van registers in artikelen in *Mens en Maatschappij* naar jaargang



Oorzaken van de verschuiving in het gebruik van enquêtes naar registers

Voor de toename van het gebruik van registers is een aantal oorzaken aan te wijzen.

De kwaliteit van de uitkomsten uit enquêtes is minder goed geworden. De kwaliteit van uitkomsten uit enquêtes hangt onder meer af van wie er wel en wie er niet meedoen aan de enquête. Als we een enquête zouden houden om het verband tussen eetgewoontes en het optreden van hart- en vaatziektes vast te stellen, en de mensen die ongezond eten maar geen last hebben van hart- en vaatziektes minder vaak meedoen, beïnvloedt dat het resultaat: het verband wordt dan overschat. In dit geval is de respons selectief.

In de loop van de jaren tachtig en negentig nam de bereidheid om aan enquêtes deel te nemen flink af. Dat gold zowel voor bedrijven en instellingen als voor personen en huishoudens. Door de forse administratieve lastendruk van het bedrijfsleven daalde de bereidheid om aan de overheid, veelal op papier, meerdere keren dezelfde gegevens te verstrekken (CBS, 1995). Ook de respons bij persoonsenquêtes is sinds de jaren tachtig in Nederland flink afgenomen. Nederland heeft gewoonlijk ook een lagere respons dan de ons omringende landen (De Leeuw & De Heer, 2002). De non-respons als zodanig is vervelend, maar leidt niet direct tot vertekening. Dat doet selectiviteit van de respons wel. De laatste jaren is er Europees gezien een tendens dat de totale respons daalt en dat dit nog sneller daalt bij specifieke moeilijk waarneembare groepen zoals allochtonen en mensen die in instellingen en tehuizen wonen (Grünwald & Körner, 2005, blz. 751). Die selectiviteit kan ook niet altijd goed gerepareerd worden door de respons te wegen (Stoop, 2005, blz. 119-123).

Tegelijkertijd nam door de ontwikkelingen in de informatietechnologie de beschikbaarheid en bruikbaarheid van registers toe. Registers zijn weliswaar voor administratieve doelen opgezet, na bewerking kunnen ze goed bruikbaar zijn voor het samenstellen van statistieken. Een kenmerk van registers is dat ze meestal een beperkte set van variabelen bevatten: in de regel de variabelen die nodig zijn voor het uitvoeren van de administratieve taken. Hierdoor is het aantal verbanden tussen variabelen dat binnen een register gelegd kan worden beperkt. Door de ontwikkelingen in de informatietechnologie kunnen grote bestanden onderling worden gekoppeld, waardoor in de resulterende bestanden deze

verbanden wel kunnen worden geschat. De invoering en steeds breder gebruik van het Sofinummer, inmiddels herdoopt tot Burger Service Nummer, heeft de mogelijkheden om registers onderling te koppelen vergroot. Omdat het CBS ook het wettelijke recht heeft verkregen van kosteloze toegang tot de administratieve gegevens die van overheidswege worden bijgehouden, zijn de mogelijkheden verder verruimd (Nobel, 2003). Het CBS stelt deze gegevens vervolgens weer ter beschikking aan sociaalwetenschappelijk onderzoekers. De gegevens zijn gratis; alleen de kosten voor het beschikbaar maken worden in rekening gebracht.

De budgettaire druk op organisaties als het CBS en de universiteiten is toegenomen. Aangezien enquêtering een duur middel is om gegevens te verzamelen, verdween dat steeds meer uit de gratie. Het gebruiken van registerinformatie voor statistische doeleinden is in de regel goedkoper dan enquêtering, omdat de registergegevens al verzameld zijn voor een administratief doel. Overigens is de kostenverlaging niet zo groot als eerst werd gedacht. Omdat de registergegevens niet zijn verzameld voor statistisch onderzoek, moeten er veel bewerkingen op plaatsvinden die niet nodig zijn bij enquêtes.

Nieuwe onderzoeksmogelijkheden

Een heel belangrijke oorzaak voor het overstappen van enquêtes op registers zijn de nieuwe onderzoeksmogelijkheden die registers bieden. Ik noem de vijf belangrijkste daarvan (zie ook Borghans, 2007):

Kleine subpopulaties

Ten eerste zijn de mogelijkheden verruimd om kleine groepen in de samenleving te onderzoeken. Alle informatie is in beginsel integraal, dat wil zeggen dekkend voor alle mensen die in Nederland wonen. Daar kun je kleine groepen uit selecteren en die aan een onderzoek onderwerpen.

Een mooi voorbeeld van een dergelijk onderzoek vind ik dat van Zorlu en Mulder (2008). Om het proces van verkleuring van wijken en buurten beter te begrijpen, hebben zij het verhuisgedrag van nieuwe immigranten onderzocht. Daarvoor gebruikten zij het Sociaal Statistisch Bestand (SSB) van het CBS. In dat SSB is een groot aantal registers koppelbaar opgenomen en voor een groot deel is de informatie onderling consistent gemaakt (Bakker, 2009). Zorlu en Mulder kozen de immigranten die in 1999 in Nederland zijn gekomen en hebben hun verhuisgedrag gedurende

vier jaar gevolgd. Daarbij hebben zij onder meer gekeken naar de invloed op het verhuisgedrag van het land van herkomst, waaronder Afghanistan en Irak, wat het migratiemotief was, dus of men voor arbeid, asiel, gezinsvorming, gezinshereniging, studie of andere redenen kwam, naar leeftijd en geslacht, of men werk heeft of niet, wat de samenstelling is van het huishouden, of men een huis huurt of niet, naar een groot aantal kenmerken van de buurt en de gemeente, allemaal op basis van registers. Het resultaat was dat het verhuisgedrag van nieuwe immigranten inderdaad bijdraagt aan de segregatie.

Als je deze studie op basis van een enquête had willen doen, dan was dat erg duur geworden. Alleen als je de beschikking zou hebben over een steekproefkader waarin de immigranten uit een bepaald jaar aangewezen kunnen worden, is het uit te voeren. Maar dan nog speelt het probleem dat het moeilijk is om deze categorie personen te benaderen, een vraaggesprek met ze te voeren en vier jaar lang te volgen. Door het onderzoek met registerinformatie te doen, kon er voor heel kleine categorieën personen uitspraken worden gedaan, zoals Irakezen en Afghanen die zich in 1999 in Nederland vestigden.

Meer dwarsverbanden en flexibiliteit

Een andere nieuwe mogelijkheid is het leggen van dwarsverbanden tussen thema's die normaal gesproken alleen via een hoge belasting voor respondenten en hoge kosten via enquêtering kunnen worden waargenomen. Daarbij komt dat actuele maatschappelijke vragen die voor hun beantwoording informatie nodig hebben die niet in een bestaande enquête voorkomt, vaak lang op zich laat wachten vanwege de productietijd van de enquête. Door systematisch registerinformatie koppelbaar beschikbaar te maken, kan een deel van de vragen die door actuele maatschappelijke ontwikkelingen opkomen snel worden beantwoord.

Een mooi voorbeeld van een dergelijke actuele vraag is de samenhang tussen voortijdig schoolverlaten en criminaliteit. Omdat de informatie over voortijdig schoolverlaten en verdenking van criminaliteit regulier in registers koppelbaar wordt opgeslagen, is het leggen van deze relatie, rekening houdend met bijvoorbeeld kenmerken van het milieu van herkomst, relatief snel en eenvoudig te doen. Deze verzameling is te vinden in het SSB.

Goedkoper en beter longitudinaal onderzoek

De volgende nieuwe onderzoeksmogelijkheid die ik wil noemen is het zonder veel extra kosten doen van longitudinaal onderzoek door personen of objecten administratief te volgen.

Een mooi voorbeeld vind ik het onderzoek dat Scheele, Van Gaalen en Van Rooijen (2009) uitvoerden naar de individuele gevolgen van massaontslag. Dat is in een periode van economische crisis natuurlijk uiterst relevante informatie. In hun onderzoek bepaalden zij eerst wie er het slachtoffer van massaontslag was in 2002, waarna deze mensen drie jaar lang in hun positie op de arbeidsmarkt werden gevolgd. Zo kon worden vastgesteld dat mensen die ontslagen waren en na drie maanden geen baan hadden, een grote kans liepen om helemaal niet meer aan de slag te komen in de drie onderzochte jaren. Verder bleek dat de mensen die een relatief slechte arbeidsmarktpositie hebben, zoals alleenstaande moeders en allochtonen, ook bij massaontslag een relatief groot risico lopen om niet meer aan de slag te komen. Verder was opvallend dat de mensen die wel snel een baan vonden, vaak een hoger loon kregen dan voordat ze het slachtoffer werden van massaontslag. Ook in dit geval werd gebruik gemaakt van louter administratieve informatie.

Het onderzoek was ook mogelijk geweest door een enquête die in het verleden is gehouden en waarin gevraagd werd of men ontslagen is en waarom, te verrijken met administratieve informatie van de jaren die volgen op de enquête. Daarmee kan het onderzoek worden uitgevoerd vrijwel zonder extra kosten en zonder de mensen persoonlijk lastig te vallen. Natuurlijk moet wel ooit een dergelijke enquête gehouden zijn, moeten de gegevens gekoppeld kunnen worden en moeten de mensen toestemming hebben gegeven daarvoor.

Het veel lastiger en duurder alternatief is een enquête in paneldesign, waarin mensen meerdere keren worden bezocht, meerdere keren worden gebeld of meerdere keren via Internet vragen beantwoorden. Werken via deze designs zijn methodologisch lastig omdat de respons in de loop der tijd afneemt, de zogenaamde panelsterfte. Het is bovendien duur omdat je vaak moet enquêteren. Een uitzondering op het laatste vormt het enquêteren via Internet. Daar zitten echter weer andere methodologische problemen aan vast. De belangrijkste daarvan is dat lang niet alle mensen toegang tot Internet hebben, waardoor de respons selectief is.

Moeilijk in enquêtes te meten onderwerpen

Een andere nieuwe mogelijkheid is onderzoek naar onderwerpen waarvan de gegevens uiterst moeizaam en met grote meetfouten in enquêtes waargenomen kunnen worden. Daarbij gaat het bijvoorbeeld om vormen van afwijkend gedrag, zoals criminaliteit of het vóórkomen van specifieke ziektes. Als een interviewer bijvoorbeeld aan u vraagt of u het afgelopen jaar wel eens crimineel bent geweest, dan verwacht ik veel vaker een eerlijk antwoord als u niet dan als u wel crimineel bent geweest.

Een voorbeeld van onderzoek met registers waarin dergelijk lastig te meten onderwerpen aan de orde komen, is de publicatie over het verband tussen herkomstgroep en criminaliteit dat ik samen met collega's van het CBS en het WODC heb geschreven (Blom et al., 2005). In die tijd was er nog volop discussie over de omvang van de criminaliteit van allochtonen en in het bijzonder van de tweede generatie. Op basis van het SSB is een verband gelegd tussen de informatie over de herkomstgroep, gebaseerd op het geboorteland van de persoon en diens vader en moeder, en het verdacht zijn van een misdrijf zoals vastgelegd in het Herkenningssysteem van de politie. Op deze wijze kon worden aangetoond dat Marokkanen van de tweede generatie, en Antillianen en Joegoslaven van de eerste generatie relatief vaak verdacht werden van een misdrijf. Ook toonden we aan dat een belangrijk deel van het verband tussen herkomstgroep en verdachte zijn wordt verklaard door demografische en sociaaleconomische kenmerken, maar dat ook na correctie nog steeds een fors verschil overblijft.

Overigens bleek uit dit onderzoek ook duidelijk dat er beperkingen kleven aan het gebruik van registers. Zo kon een fors deel van de verdachten niet worden gekoppeld, omdat ze niet voldoende geïdentificeerd konden worden. Het is duidelijk dat verdachten er ook alle belang bij hebben om niet geïdentificeerd te worden. Daarnaast is verdacht zijn van een misdrijf iets anders dan dader zijn, al wordt ongeveer 90% van de verdachten uiteindelijk ook veroordeeld.

Toegenomen mogelijkheden voor intergenerationeel onderzoek

De laatste nieuwe onderzoeksmogelijkheid die ik noem is het leggen van verbanden tussen de kenmerken van ouders en kinderen, die ook weer alleen via een grote inspanning via enquêtering verkregen kunnen worden. Omdat we in de Gemeentelijke Basisadministratie persoonsgegevens

(GBA) veel relaties tussen ouders en kinderen kunnen leggen, en het aantal relaties dat gelegd kan worden bovendien ieder jaar toeneemt, kunnen de administratieve gegevens van ouders en kinderen worden gecombineerd.

Een goed voorbeeld van intergenerationeel onderzoek dat je alleen met een grote inspanning via enquêtes zou kunnen doen, is het onderzoek naar het verband tussen het crimineel gedrag van ouders en kinderen dat is uitgevoerd door Besjes en Van Gaalen (2008). Daarin wordt informatie gebruikt over de criminele carrières van vaders en moeders, en die van hun zonen en dochters. Bovendien wordt de samenhang met allerlei demografische en sociaaleconomische kenmerken van ouders en kinderen in het onderzoek betrokken. Het onderzoek toont aan dat volwassen kinderen met een ouder die verdacht wordt van criminaliteit ook zelf een grotere kans lopen crimineel te zijn. Bovendien blijkt dat een criminele moeder wat dit betreft nog grotere risico's oplevert dan een criminele vader.

Als dit soort onderzoek wordt gedaan met behulp van een enquête, zou allereerst een goed steekproefkader moeten worden gezocht voor de mensen die crimineel zijn. Vervolgens zouden die mensen òf naar de criminele carrière van hun ouders moeten worden gevraagd, òf de ouders zouden afzonderlijk moeten worden bevraagd. Een ingewikkeld en duur onderzoeksdesign.

Ik ga hier geen lofzang houden over het onderzoek met behulp van registers, maar het biedt zeker nieuwe mogelijkheden. Trek alle registers open, zou ik zeggen, en benut die mogelijkheden. Ik wil ook niet zeggen dat enquêtering niet langer nodig is. Een deel van de voor officiële overheidsstatistieken en sociaalwetenschappelijk onderzoek benodigde informatie kan immers niet aan registers worden ontleend. Daarbij gaat het bijvoorbeeld om meningen en houdingen, of om feitelijk gedrag dat niet geregistreerd wordt. Voor het beantwoorden van verklarende sociaalwetenschappelijke vragen zal ook vaak een combinatie van enquête- en registerinformatie nodig zijn, omdat dikwijls de verklarende variabelen niet voldoende in registers zijn opgenomen.

De onderzoeksvraag: de kwaliteit van registers

Hoewel registers steeds meer gebruikt worden, is over mogelijke problemen met de kwaliteit ervan vrijwel geen literatuur hoewel evident is

dat ook registers kwaliteitsproblemen hebben (Grünewald & Körner, 2005, blz. 752-753). De kwaliteit is echter vaak onbekend. De registerhouder heeft er alle belang bij om eventuele kwaliteitsproblemen niet aan de grote klok te hangen. De kwaliteit van een register heeft dikwijls direct te maken met het functioneren van de registerhouder. Een gebrekkige kwaliteit documenteren en openbaar maken zal vaak worden beschouwd als “het buiten hangen van de vuile was”. Een bijkomend probleem hierbij is dat bij registerhouders meestal geen statistici werken die weten hoe je de kwaliteit systematisch in kaart moet brengen. Dit laatste is een meer algemeen probleem. Omdat statistici meestal zijn opgeleid in een periode dat statistisch onderzoek vrijwel uitsluitend op basis van enquêtes werd uitgevoerd en ook hun arbeidservaring zich meestal daartoe beperkt, schiet hun kennis hierover vaak ook te kort.

De vraag is welke kwaliteitsproblemen van registers in beeld gebracht zouden moeten worden. Het onderkennen van deze problemen, het ontwikkelen van methodes om deze op het spoor te komen, de omvang ervan te schatten en methoden ontwikkelen om de problemen op te lossen is de uitdaging waar we voor staan. Met andere woorden: welke informatie moeten we opnemen in de paragraaf van een wetenschappelijk artikel als daarin gebruik wordt gemaakt van registers? En welke methodes lenen zich voor het verwerken en analyseren van registerdata?

Uitgangspunt voor de documentatie van de kwaliteit van registers

Wat is nu eigenlijk een register? Een register wordt gedefinieerd als een volledige lijst van de objecten die deel uitmaken van een (sub)populatie en de bij deze objecten behorende kenmerken aan te duiden door een waarde op een variabele. Objecten kunnen echter worden gemist vanwege kwaliteitsproblemen in het register. Registers hoeven daarom dus niet volledig te zijn. Om de gegevens in een onderzoeksomgeving te kunnen gebruiken is het noodzakelijk dat de objecten eenduidig kunnen worden geïdentificeerd, zodat de gegevens kunnen worden geactualiseerd en gekoppeld (Wallgren & Wallgren, 2007, blz. 4).

Het uitgangspunt voor de beschrijving van de kwaliteit van registers is dat deze dezelfde soorten fouten bevatten als enquêtes. Hoewel het registreren voor andere doeleinden wordt gedaan dan voor het verzamelen van statistische gegevens, kan men het registratieproces als zodanig wel

beschouwen als een soort enquête. Vaak moeten (al dan niet elektronische) formulieren worden ingevuld bij registratie of wordt door de registerhouder informatie op grond van een vraaggesprek ingevoerd in een database.

Laat ik enkele voorbeelden noemen om te illustreren dat registers vaak gevuld worden via een enquête. Eens per jaar vullen de meesten van u een aangifteformulier in voor de inkomstenbelasting. Dat kan elektronisch of dat kan op papier. Het elektronisch invullen van uw belastingformulier is een vorm van elektronische enquêtering, qua methodologie vergelijkbaar met het zogenaamde Computer Assisted Web Interviewing (CAWI). Als u het papieren formulier invult lijkt dat veel op Paper and Pencil Interviewing (PAPI). De Belastingdienst besteedt veel aandacht aan het ontwerp van deze elektronische en papieren formulieren en zorgt ook voor de vaak uitgebreide toelichtingen op papier of in de helpfunctie. Deze toelichtingen zijn pogingen om de interpretatie van de gestelde vragen zodanig te beïnvloeden dat daarover geen misverstanden bestaan.

Een tweede voorbeeld uit de praktijk is als u zich moet inschrijven bij het UWV als werkzoekende. Daarbij is sprake van enquêtering in twee rondes. In de eerste ronde wordt de werkzoekende geacht om een elektronische vragenlijst via Internet in te vullen. Daarin worden bijvoorbeeld vragen gesteld over het arbeidsverleden, over de opleidingen die de werkzoekende heeft gevolgd en afgerond en de wensen die de werkzoekende heeft over het nieuwe werk en de nieuwe werkgever. Dit is weer een vorm van CAWI. De enquêtering in de eerste ronde dient als uitgangspunt voor de tweede ronde waarin de arbeidsbemiddelaar op basis van een gesprek met de werkzoekende een elektronische database vult. De volgorde waarin deze database gevuld wordt en de formulering van de vragen wordt aan de arbeidsbemiddelaar overgelaten, zij het dat de invulschermen daarin wel sturend zijn. Dit is een vorm van half gestructureerd Computer Assisted Personal Interviewing (CAPI).

Overigens zijn er registers waarin de rol van enquêtering wat minder groot lijkt te zijn. Zo worden veel gegevens ontleend aan de zogenaamde Polisadministratie van het UWV die grotendeels wordt gevuld door bedrijven voor hun werknemers, door pensioenfondsen voor hun gepensioneerden en door uitkeringsinstanties voor hun uitkeringsgerechtigden. Er zijn weliswaar wel wat elementen van enquêtes aan te pas gekomen om de registers te vullen, maar het zijn toch vooral de

administratieve afdelingen van de bedrijven, pensioenfondsen en uitkeringsinstanties die de uiteindelijke inhoud bepalen. Dan zijn de administratieve processen veel bepalender voor de kwaliteit van de informatie die aan dergelijke registers kan worden ontleend.

De mogelijke foutenbronnen in onderzoek op basis van enquêtes

Om een overzicht te geven van de mogelijke fouten in empirisch onderzoek waarin registers worden gebruikt, kan de constatering dat het proces van registratie elementen bevat van enquêtes goede diensten bewijzen. In ieder geval zullen de fouten die in enquêtes voorkomen ook in registeronderzoek voorkomen. Als uitgangspunt van mogelijke fouten in registeronderzoek gebruik ik het overzicht zoals Groves *et al.* (2004, blz. 39-65) dat schetsen. Zij beschrijven de ‘totale enquêtefout’ en leggen deze uiteen in verschillende componenten. Daarbij wordt door hen de ‘levensloop’ van een enquête gevolgd en wordt een onderscheid gemaakt tussen meting- en representatiefouten (figuur 2). De fouten die in deze figuur zijn opgesomd zijn eigenschappen van een enkele uitkomst van een enquête en niet van de enquête als geheel. Het gaat om het schatten van een verband tussen twee variabelen. De fouten in de meting gelden voor beide variabelen. Daarom is van iedere ovaal ook een schaduw aangegeven voor de tweede variabele. De genoemde representatiefouten zijn op alle uitkomsten van een enquête van toepassing; de omvang van die fouten kan wel per uitkomst verschillen.

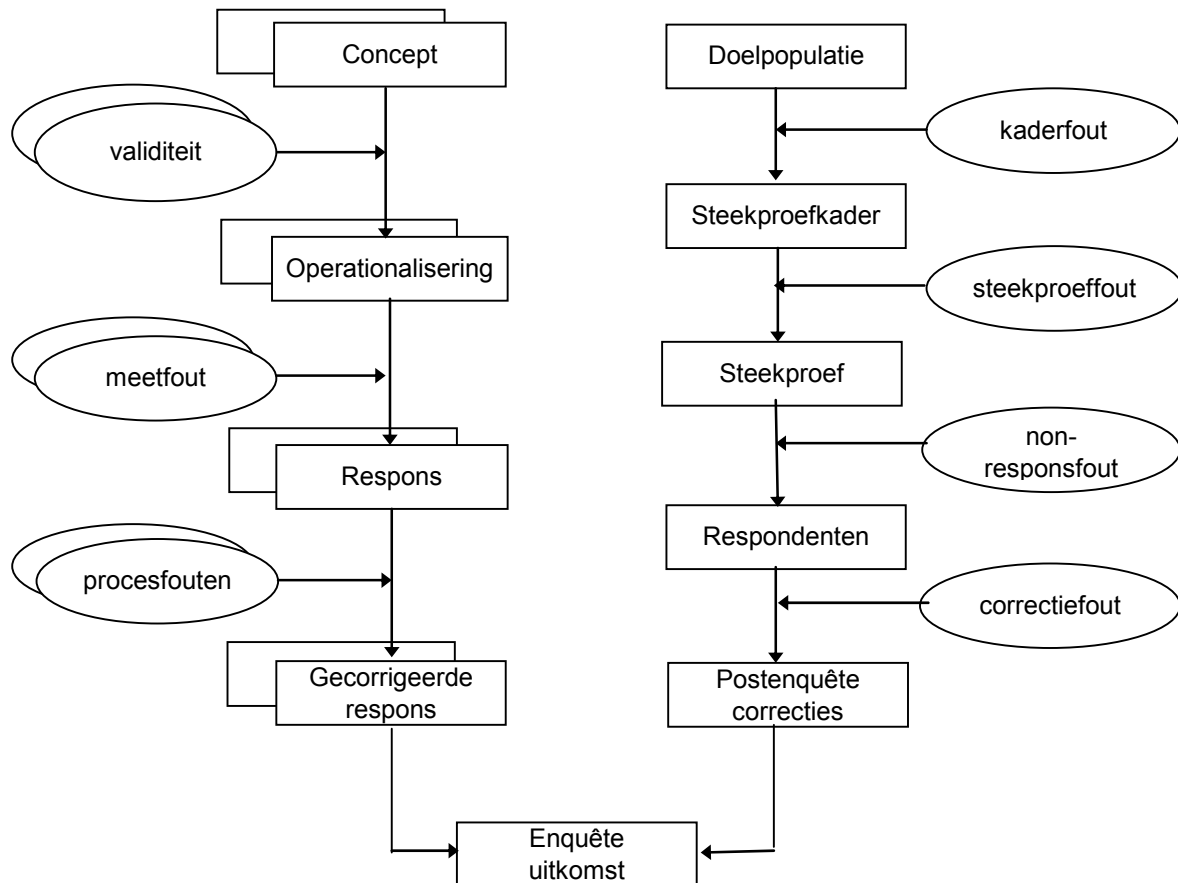
Aan de linkerkant van de figuur, aan de kant van de meting dus, starten we met het concept van het te meten begrip. Dat kan bijvoorbeeld het opleidingsniveau zijn. Het begrip wordt geoperationaliseerd in een vraagstelling. Bijvoorbeeld de vraag “Wat is uw hoogst genoten opleiding?”, waarbij een aantal antwoordcategorieën wordt voorgelegd. Bij de operationalisering kun je fouten maken die zorgen voor systematische vertekening in de uitkomsten. Zo weten we dat in het bijzonder de voorgelegde antwoordcategorieën van belang zijn hiervoor.

Vervolgens wordt daadwerkelijk de vragenlijst afgenomen. Daarbij worden meetfouten gemaakt. Ik noem er een paar. In de eerste plaats ontstaan er fouten omdat mensen zich bepaalde zaken niet meer goed kunnen herinneren. Als ik aan u vraag of u de afgelopen drie maanden de huisarts heeft bezocht, dan kunt u een eventueel bezoek in deze periode zijn

vergeten of u kunt ten onrechte denken dat het langer geleden is dan drie maanden. Andersom kan het ook zijn dat u een huisartsbezoek van langer dan drie maanden geleden in de referentieperiode trekt (Sikkel, 1988).

Figuur 2.

De ‘levensloop’ en de bronnen van fouten in een enquête



In de tweede plaats kan ook de interviewer specifieke meetfouten veroorzaken (Pannekoek, 1988; Ganninger, Häder & Gabler, 2007; Brick et al., 200). Door vragen op een bepaalde manier te stellen worden ook bepaalde antwoorden uitgelokt. Suggestieve vragen van de vorm “U wilt zeker geen kinderen meer” zijn daarvan een goed voorbeeld. Een derde fout is sociaal wenselijk antwoordgedrag (Belson, 1986) dat optreedt als de geïnterviewde het vervelend vindt om naar waarheid te antwoorden omdat dat ongepast is. Een goed voorbeeld daarvan is het niet rapporteren van eigen crimineel gedrag.

Het afnemen van de vragenlijst leidt tot een respons. Nadat de vraaggesprekken hebben plaatsgevonden, worden de gegevens gecorrigeerd met behulp van een stelsel van beslisregels. Daarbij worden onwaarschijnlijke waarden geschrapt en soms vervangen, worden onbeantwoorde vragen alsnog geschat met behulp van een model, worden nieuwe variabelen afgeleid door de antwoorden van verschillende vragen te combineren in bijvoorbeeld schalen of door alfanumerieke informatie te coderen. Bij al deze verwerkingsstappen kunnen foute beslisregels worden geformuleerd en toegepast. Deze dragen bij aan de totale enquêtefout. Na de verwerking van de gegevens spreken we over gecorrigeerde respons.

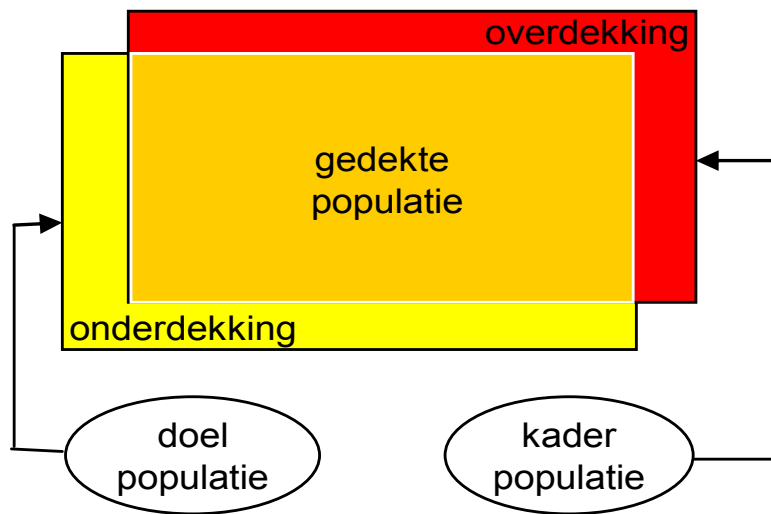
Aan de rechterkant van figuur 2 wordt de representatie van de enquêteresultaten weergegeven. Hier gaat het om de mate waarin de enquêteresultaten een goede afspiegeling zijn van de populatie waarover je uitspraken wilt doen. Daarbij wordt begonnen met het definiëren van de populatie, bijvoorbeeld de bevolking van Nederland van 15 jaar of ouder op 1 november 2009.

Voor deze doelpopulatie wordt een steekproefkader gezocht dat idealiter alle elementen van deze doelpopulatie bevat. Het is echter niet altijd mogelijk om een steekproefkader te vinden of op te bouwen waarin exact alle elementen uit de doelpopulatie zijn opgenomen. Dit kan leiden tot onderdekking of overdekking (zie figuur 3). Van onderdekking is sprake als er elementen van de doelpopulatie niet in het steekproefkader opgenomen zijn. Van overdekking is sprake als er elementen wel in het steekproefkader opgenomen zijn die geen deel uitmaken van de doelpopulatie.

Een voorbeeld kan dat verduidelijken. Voor de zojuist gedefinieerde doelpopulatie van de bevolking van Nederland van 15 jaar en ouder op 1 november 2009 kiezen we als steekproefkader de bevolking die op 1 november 2009 ingeschreven is in de GBA en op dat moment 15 jaar of ouder was. Deze keuze leidt tot kaderfouten. Er is sprake van onderdekking omdat in het kader bijvoorbeeld geen mensen die illegaal in Nederland verblijven opgenomen zijn, en er is sprake van overdekking omdat er bijvoorbeeld mensen zijn geëmigreerd vóór 1 november maar dat nog niet hebben doorgegeven aan de gemeente waaruit ze vertrokken.

Figuur 3.

Over- en onderdekking door kaderfouten bij enquêtes



Uit het steekproefkader wordt voor een enquête een steekproef getrokken en vervolgens worden aan de getrokken steekproefelementen vragenlijsten voorgelegd. Als er meerdere steekproeven getrokken worden en de elementen steeds ondervraagd zouden worden, zou er spreiding in de uitkomsten ontstaan. Deze uitkomsten liggen, indien er geen andere fouten worden gemaakt, rond de werkelijke waarde. Deze spreiding noemen we de steekproefvariantie. Als één steekproef wordt getrokken kun je een toevalstreffer hebben waarvan de waarde van de uitkomsten ver afligt van de werkelijke waarde. Dat noemen we de steekproeffout.

Nadat de steekproefelementen bepaald zijn, worden deze benaderd voor een interview. In de praktijk doen lang niet alle mensen die benaderd worden ook daadwerkelijk mee. Deze non-respons leidt tot twee fouten. In de eerste plaats leidt het tot minder waarnemingen waardoor de variantie hoger is dan bij volledige waarneming. Dat is echter niet zo erg. In de tweede plaats kan het tot vertekening leiden als de non-respons niet toevallig is verdeeld over de kenmerken waarnaar onderzoek wordt verricht. Als de non-respondenten op deze kenmerken afwijken van de respondenten dan zijn de uitkomsten vertekend. Denk nog maar eens aan het voorbeeld dat ik eerder genoemd heb over het verband tussen eetgewoontes en het voorkomen van hart- en vaatziekte.

De laatste stap in het gehele proces bestaat eruit dat de kader-, steekproef- en non-responsfouten zoveel mogelijk worden gecorrigeerd. Bij deze stap wordt meestal informatie gebruikt die bekend is voor de doelpopulatie uit registers of uit het steekproefkader. De categorieën respondenten die zijn ondervertegenwoordigd krijgen een hoger gewicht en de categorieën respondenten die zijn oververtegenwoordigd krijgen een lager gewicht. Op deze manier worden de uitkomsten voor de respons gelijk gemaakt aan die uit de doelpopulatie. Dat geldt overigens alleen voor de variabelen die betrokken worden in de weging en geeft geen garantie voor de uitkomsten van andere variabelen. De fouten die bij dit correctieproces worden gemaakt noemen we correctiefouten.

De mogelijke foutenbronnen in onderzoek op basis van registers

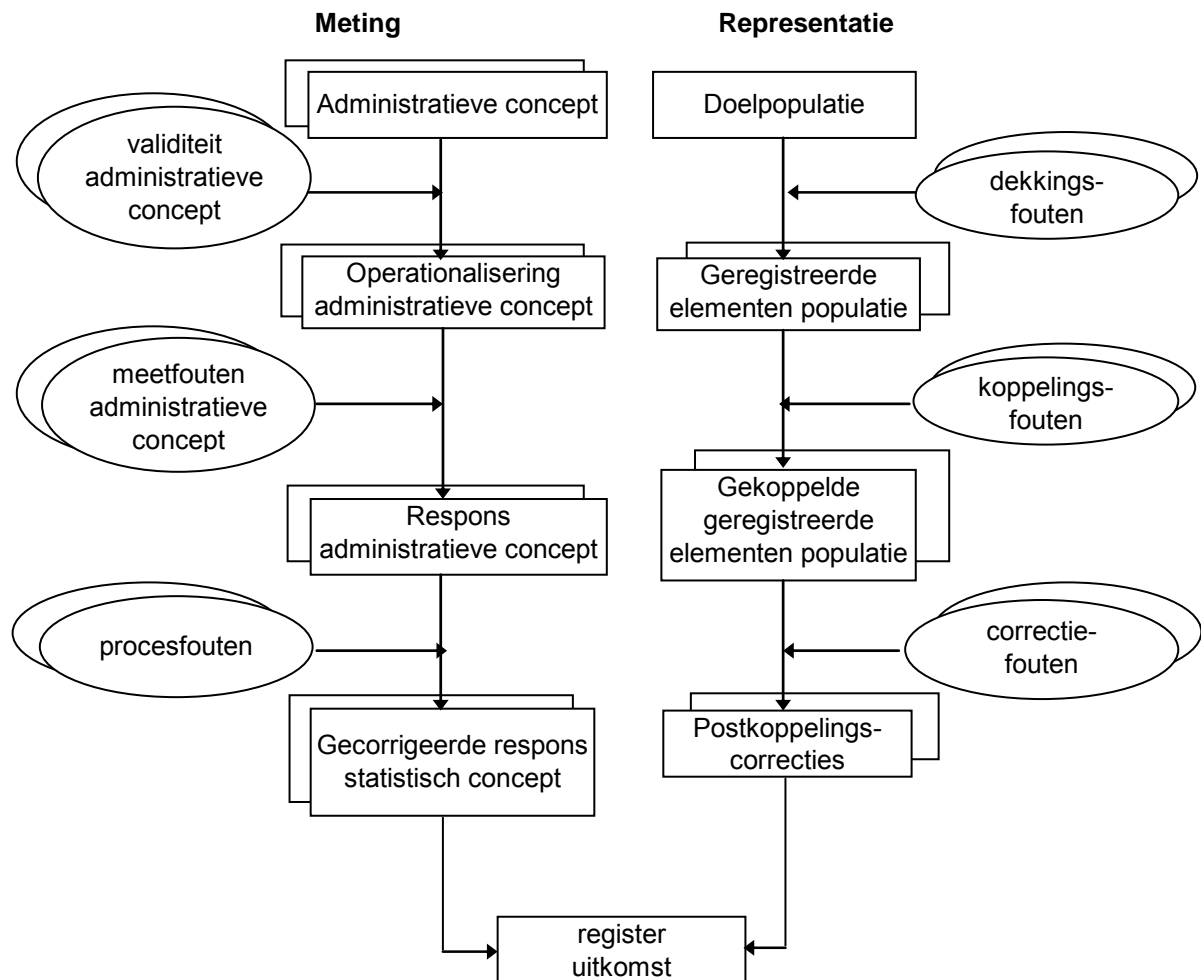
De mogelijke fouten bij enquêtes komen ook voor in registers. Dat wil niet zeggen dat de fouten identiek zijn. Er zijn verschillen in omvang, er zijn verschillen in de mate waarin de fouten kunnen worden beïnvloed en er zijn voor registers unieke fouten (zie ook Wallgren en Wallgren, 2007). In figuur 4 heb ik de mogelijke fouten in registers weergegeven. De uitkomst van het proces is de schatting van het verband tussen twee variabelen uit verschillende registers, vandaar dat er twee ovalen en twee rechthoeken in de figuur zijn opgenomen voor ieder van de stappen in de levensloop van de registers.

De *omvang van de fouten in registers* wordt mede bepaald door de mate waarin fouten door de registerhouder worden gesignaleerd en gecorrigeerd. Uiteraard worden fouten in registers zoveel mogelijk voorkomen door het inbouwen van mechanismen in het administratieve proces om mogelijke fouten te voorkomen of te corrigeren. Zo kunnen arbeidsbemiddelaars bij het vaststellen van het opleidingsniveau vragen naar de diploma's die de cliënt behaald heeft. Ook worden er andere administratieve bestanden gekoppeld om te zien of de informatie consistent is. Deze administratieve protocollen richten zich op de variabelen die van belang zijn voor de registerhouder. Dit gezegd hebbende kunnen we ook veronderstellen dat de omvang van de fout mede afhangt van het belang dat de registerhouder heeft bij het juist vastleggen van de informatie. Een voorbeeld van informatie die bij de aangifte van de inkomstenbelasting niet van belang is voor de Belastingdienst, is de beëindigingsdatum van een dienstverband. De afdracht van premies en inkomstenbelasting is daarvan niet afhankelijk.

Het blijkt in de praktijk dan ook een variabele van relatief lage kwaliteit op te leveren.

Figuur 4

De ‘levensloop’ en de bronnen van fouten in gekoppelde registers



De *invloed* die een onderzoeker kan uitoefenen *op het gemeten concept* is bij registers kleiner dan bij enquêtes. De registers zijn immers opgezet voor het vastleggen van administratieve concepten. Het afleiden van de statistische concepten van de administratieve data noemen we harmonisatie. Als het administratieve concept afwijkt van het statistische concept, kun je daar in de regel maar weinig aan veranderen. Soms kun je door verschillende administratieve variabelen te combineren statistische concepten beter benaderen dan met een enkele administratieve variabele, maar dikwijls ontbreekt deze mogelijkheid. Wat dan rest is te werken met

het administratieve concept en je bewust te zijn van de beperkingen daarvan.

Administratieve vertraging is een belangrijke foutenbron in registers. Deze treedt op wanneer een gebeurtenis op een veel later tijdstip in het register wordt verwerkt dan waarop het heeft plaatsgevonden. Het meest pregnante voorbeeld daarvan is de vertraging in de registratie van huwelijken van allochtonen die trouwen in hun land van herkomst. Vaak wordt pas als de huwelijkspartner daadwerkelijk overkomt het huwelijk in de GBA geregistreerd. Soms gaat daar een periode van 1 à 2 jaar overheen (zie ook Hooghiemstra, 2003). Meestal zijn de administratieve vertragingen minder groot. Andere voorbeelden zijn regelingen waarin zaken met terugwerkende kracht worden ingevoerd, zoals de toekenningen van het recht op een arbeidsongeschiktheid- of een bijstanduitkering. Daarbij wordt dat recht vaak een paar maanden na het ontstaan van het recht toegekend. Administratieve vertraging leidt tot onderschatting van het aantal gebeurtenissen in een bepaalde referentieperiode. De manier om dit probleem op te lossen is net zolang te wachten met publiceren tot alle vertraagde informatie beschikbaar is. Dit is echter onverenigbaar met de wens om actuele informatie over maatschappelijke ontwikkelingen te publiceren. Daarom moeten we onze toevlucht nemen tot het schatten van de omvang van de administratieve vertraging en de actuele cijfers daarvoor corrigeren.

Wat we vaak tegenkomen is dat inconsistenties worden opgelost met een beslisregel. Dit leidt voor het overgrote deel van de objecten tot een juiste uitkomst, maar betekent voor een klein deel van de objecten dat er fouten worden gecreëerd. Dit zou betekenen dat het corrigeren met behulp van beslisregels leidt tot regressie het gemiddelde. Een alternatief hiervoor is om, als de omvang van de fout kan worden berekend, met een kansverdeling de beslisregel op een deel van de desbetreffende (sub)populatie toe te passen. Er vindt dan geen regressie naar het gemiddelde plaats, maar het aantal fouten op individueel niveau is wel groter (Bakker, 2009).

Aan de representatiekant kan het zijn dat de bronnen die gecombineerd worden geen van alle volledig de doelpopulatie dekken. Voor ieder van de bronnen kan gelden dat zij zowel over- als onderdekking kennen. Deze *dekkingsfouten* worden zoveel mogelijk opgelost door te bepalen welke elementen tot de doelpopulatie behoren en alle elementen uit de bronnen

die daaraan voldoen samen te voegen. Dat noemen we completeren. Uiteraard kunnen niet alle dekkingsfouten door completering opgelost worden als in de gecombineerde bronnen structureel een deel van de doelpopulatie ontbreekt (Bakker, 2009).

Een ander kwaliteitsprobleem kan optreden als de *koppeling van registers* teveel gemiste of miskoppelingen oplevert. Gemiste koppelingen kunnen leiden tot selectiviteit die vergelijkbaar is met selectiviteit veroorzaakt door non-respons. Uit eerder onderzoek blijkt dat de gebruikelijke koppelingsprocedures tot enige selectiviteit leiden (Arts, Bakker en Van Lith, 2000), zij het dat de selectiviteit kleiner is dan die veroorzaakt door non-respons bij enquêtes. Er kan dan op dezelfde wijze voor worden gecorrigeerd, namelijk door weging (vgl. Bakker & Van Rooijen, 2000, Wallgren & Wallgren, 2007, blz. 127-146). Daarbij geldt overigens, net als bij de weging van enquêtes, dat aangenomen wordt dat gegeven de configuratie van hulpvariabelen de gekoppelde geregistreerden een goede afspiegeling vormen van de niet-gekoppelde geregistreerden. Of dit ook zo is, is echter veel beter te onderzoeken bij registers dan bij enquêtes, omdat ook voor de niet-gekoppelde geregistreerden de informatie uit de registers beschikbaar is. Die informatie kan vergeleken worden met die van wel geregistreerden om zo de mate van selectiviteit vast te stellen.

Miskoppelingen tussen twee registers leveren geheel andere problemen op. Hiervan is sprake als de informatie van twee verschillende personen wordt gekoppeld, ondanks het feit dat zij gelijk zijn op de variabelen waarop gekoppeld wordt. Deze bron van fouten is te beschouwen als een bijzondere vorm van de fouten die gemaakt worden bij statistisch of probabilistisch koppelen van meerdere steekproeven. Dit leidt tot vertekening en meestal tot onderschatting van de sterkte van verbanden tussen variabelen (Fellegi en Sunter, 1969; Kroese, Renssen & Trijssenaar, 2000). Bovendien leidt het koppelen van de informatie van twee verschillende personen tot veel inconsistenties in de data. Bij het correctieproces heb je daar veel last van.

Op grond van het verhaal dat ik tot nu toe heb gehouden is duidelijk op welk onderzoeksterrein ik me ga begeven. Maar het is ook een theoretisch verhaal. Ik zal nu aan de hand van twee concrete voorbeelden aangeven welk onderzoek ik in de komende periode zal uitvoeren. Het eerste gaat over onderzoek naar de validiteit en betrouwbaarheid in de meting van een variabele en is dus exemplarisch voor de linkerkolom van de figuur. Het

tweede gaat over de over- en onderdekking van populaties en is een voorbeeld van fouten uit de tweede kolom. Daarbij wordt de vraag beantwoord welke over- en onderdekking er is als de GBA wordt gebruikt voor de schatting van de omvang van de meest gebruikte populatie in Nederland: de Nederlandse bevolking.

Modelleren van meetfouten

Een van de meetfouten die in registers voorkomt is vertekening, dat wil zeggen dat je niet precies dát meet wat je wilt meten. Als er sprake is van grote vertekening, zeggen we ook wel dat de validiteit van het meetinstrument laag is. Dat zijn systematische fouten. Naast de vertekening kennen metingen ook een zekere onbetrouwbaarheid, dat zijn willekeurige meetfouten. Stel dat we het ware gemiddelde bruto loon η van werknemers in Nederland zouden weten. Als we het gemiddelde zouden willen schatten met behulp van meerdere steekproeven, dan zou er op basis van toevallige fluctuaties een spreiding van geschatte gemiddelde bruto lonen ontstaan. Deze spreiding noemen we de betrouwbaarheid. Als het netjes is verdeeld rondom het ware gemiddelde, is er geen vertekening. Als het echter niet rondom het ware gemiddelde loon ligt, dan is sprake van vertekening. Tot zover de elementaire statistiek. Hoe zou je echter de betrouwbaarheid en de vertekening kunnen vaststellen?

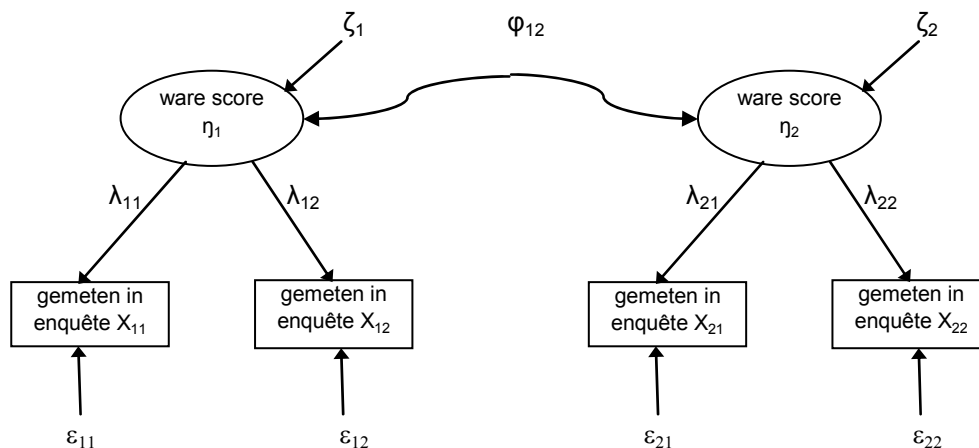
De klassieke testtheorie is ontwikkeld om de betrouwbaarheid en validiteit van meetinstrumenten te bepalen (Verschuren, 1991; Jöreskog en Sörbom, 1996; Kline, 2005). Stel dat we een enquête hebben waarin we het concept zonder vertekening hebben gemeten. We onderzoeken de validiteit van de tweede enquête waarin we hetzelfde concept hebben gemeten. De correlatie tussen die twee is dan de validiteit van de meting in de tweede enquête. Deze wordt als het ware geijkt aan de eerste.

De variabelen in beide enquêtes zijn echter ook met een zekere onbetrouwbaarheid gemeten. Om de validiteit goed te kunnen bepalen moet voor de onbetrouwbaarheid van de variabelen in beide enquêtes worden gecorrigeerd. De betrouwbaarheid van de meting van een variabele wordt vastgesteld door deze twee keer binnen een korte periode te verrichten. Stel dat ik de betrouwbaarheid van de vraagstelling naar het loon zou willen vaststellen, dan stel ik de vraag nu en over twee weken weer. De correlatie is dan een schatting van de betrouwbaarheid. Dit wordt

de test-hertestcorrelatie genoemd. Het model dat geschat wordt, is weergegeven in figuur 5.

Figuur 5

Een model voor de bepaling van de kruisvaliditeit van een variabele met behulp van twee enquêtes



Een voorbeeld kan dit verder verduidelijken. Doel is om de validiteit van de variabele persoonlijk bruto loon X_1 in de eerste enquête te bepalen. In deze enquête is daarvoor een simpele vraagstelling opgenomen. Deze enquête wordt twee keer afgenomen en leidt daarmee tot de scores X_{11} en X_{12} . De overeenkomst in de antwoorden is dan de maat voor de betrouwbaarheid. Het product van de parameters λ_{11} en λ_{12} is gelijk aan de test-hertestcorrelatie. De voor onbetrouwbaarheid gecorrigeerde waarde van de variabele noemen we de 'ware score' η_1 . Verder wordt hetzelfde kenmerk gemeten met behulp van een tweede enquête: X_2 . De vraagstelling in de tweede enquête is uitgebreid en de validiteit daarvan is al eerder vastgesteld en erg hoog bevonden. Ook hiervoor geldt dat de meting twee keer binnen een korte periode is verricht. Dit leidt tot de scores X_{21} en X_{22} . De validiteit van de meting in de eerste enquête is de 'ware' correlatie tussen η_1 en η_2 : φ_{12} . Zo kunnen we door gebruik te maken van het idee van kruisvaliditeit en de klassieke testtheorie meetfouten modelleren.

Kunnen we dezelfde methode ook toepassen voor het vaststellen van de validiteit van registervariabelen? Daarvoor heb je een enquête nodig waarin de variabele valide en herhaald wordt gemeten en, daaraan

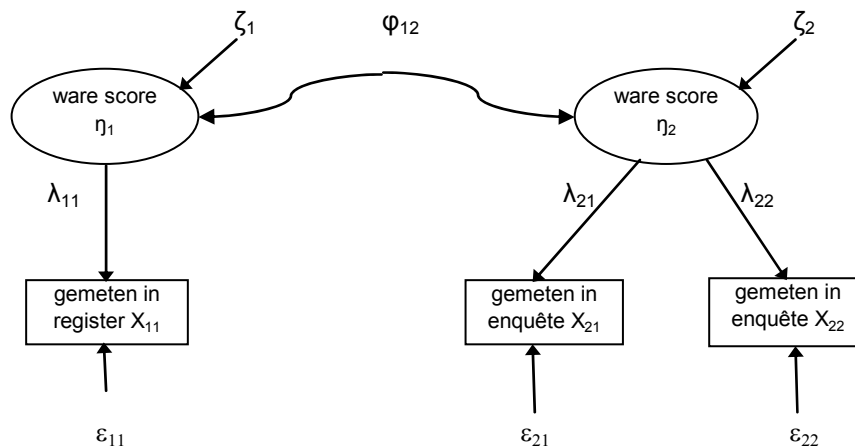
gekoppeld, een of meerdere registers waarin de variabele herhaaldelijk voorkomt. Het herhaald meten van variabelen in grootschalige enquêtes wordt slechts zelden daadwerkelijk uitgevoerd. De belangrijkste redenen daarvoor zijn dat het kostbaar is en dat het een zware belasting is van de respondenten. Daarom zal dat slechts op kleine schaal kunnen gebeuren. Voor registers is een herhaalde meting met een korte periode tussen de twee metingen helemaal niet realistisch. Die herhaalde meting betekent immers dat we het registratieproces twee keer volledig zouden doorlopen. Het is waarschijnlijk erg lastig om medewerking daarvoor te verkrijgen van registerhouders en geregistreerden. Wel kan er gebruik worden gemaakt van achtereenvolgende leveringen die betrekking hebben op aansluitende tijdsperiodes. Zo worden de gegevens over banen en de bijbehorende bruto lonen in de regel maandelijks geleverd. De metingen van twee achtereenvolgende maanden kunnen worden beschouwd als een herhaalde meting. Daar zitten echter wel enkele methodologische problemen aan vast. In de eerste plaats zijn de metingen niet onafhankelijk van elkaar, omdat ze afkomstig zijn uit dezelfde administratieve systemen. Dat kan leiden tot gecorreleerde meetfouten. Die kunnen wel in het model worden ingebouwd, al zal het lastiger zijn om de parameters te berekenen. In de tweede plaats kan er daadwerkelijk iets aan het loon zijn veranderd omdat de metingen niet betrekking hebben op dezelfde tijdsperiode. Dat zou leiden tot een onderschatting van de betrouwbaarheid.

Een andere mogelijkheid is om gegevens over dezelfde variabele uit verschillende registers die betrekking hebben op dezelfde eenheden en peilmoment te gebruiken als herhaalde meting. Er kunnen dan wel conceptuele verschillen zijn. Die kunnen dan uiteraard ook weer in het model zichtbaar worden gemaakt. Je kunt de bovenstaande twee manieren van herhaalde meting ook combineren.

Als herhaalde meting wel in een enquête wordt toegepast, maar niet in het register, wordt het model geschat zoals in figuur 6. Daarbij moet noodgedwongen worden aangenomen dat $\lambda_{11}=1$. De registermeting wordt gecorreleerd met de voor onbetrouwbaarheid gecorrigeerde meting in de enquête. Het geschatte model in figuur 6 is niet geïdentificeerd, dat wil zeggen dat er meer onbekenden dan vergelijkingen zijn, waardoor de parameters in het model niet uitgerekend kunnen worden.

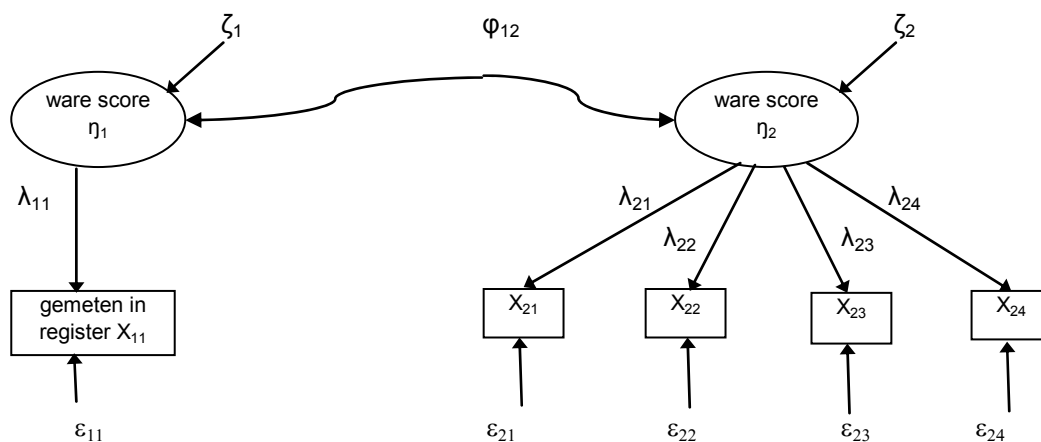
Figuur 6

Een model voor de bepaling van de kruisvaliditeit van een registervariabele met behulp van een enquêtevariabele



Figuur 7

Een model voor de bepaling van de kruisvaliditeit van een registervariabele met behulp van een schaal geconstrueerd in een enquête



Wat vaker als alternatief wordt gehanteerd, is dat meerdere vragen over het ongemeten kenmerk worden gesteld. Door daarvan schalen te construeren, kan de betrouwbaarheid worden verhoogd. Daarmee wordt het model bovendien geïdentificeerd en kunnen de parameters dus worden uitgerekend (zie figuur 7).

We gaan er bij het toepassen van de klassieke testtheorie steeds van uit dat er een bron is waarvan we vastgesteld hebben dat de validiteit hoog is. In

de praktijk is de validiteit van vraagstellingen in bestaande enquêtes echter meestal niet bekend. Het is daarom beter om een enquête exclusief te ontwerpen om de kwaliteit van registerinformatie te beoordelen. Deze enquête kan best kleinschalig zijn, als de totale enquêtetfout maar klein is. Een paar honderd waarnemingen is in de meeste gevallen voldoende. Als we uit onderzoek naar de administratieve praktijk een kwalitatief idee hebben over de fouten die in een register bij bepaalde variabelen voorkomen, kan door daarover vragen te stellen in een enquête ook de omvang daarvan worden vastgesteld. Bij het bepalen van de omvang van de fouten kunnen modellen worden gebruikt die ik zojuist heb besproken.

Andersom gebeurt overigens vaker: om de meetfouten in enquêtes te onderzoeken worden de uitkomsten van enquêtes gekoppeld aan registers. Aangenomen wordt vervolgens dat de registers geen meetfouten hebben, waarna de kwaliteit van de enquêtedata vastgesteld kan worden (zie voor voorbeelden Duncan & Hill, 1985; Rodgers, Brown & Duncan, 1993; Bound et al., 1994; Pischke, 1995; Pedace and Bates, 2001; Stinson, 2002; Kim & Solon, 2005). Daarnaast worden registers gebruikt voor het bepalen van over- en onderdekking van enquêtes en het bepalen van en corrigeren voor de selectiviteit van de non-response (Mulry et al., 2006). In sommige gevallen wordt wel de fout veroorzaakt door het koppelen gemodelleerd (Kapteyn & Ypma, 2007), maar niet de fouten veroorzaakt door de waarneming via registers.

Waarom moet je je nu zo druk maken over de betrouwbaarheid en vertekening? Voor de vertekening lijkt mij dat evident. Als je iets anders meet dan je zou moeten meten om de hypothesen te toetsen die afgeleid worden van de theorieën, schiet je je doel voorbij. Je druk maken over de betrouwbaarheid is ook noodzakelijk, omdat de betrouwbaarheid invloed heeft op de schatting van de omvang van de relaties tussen variabelen, iets wat het al dan niet verwerpen van hypothesen kan beïnvloeden.

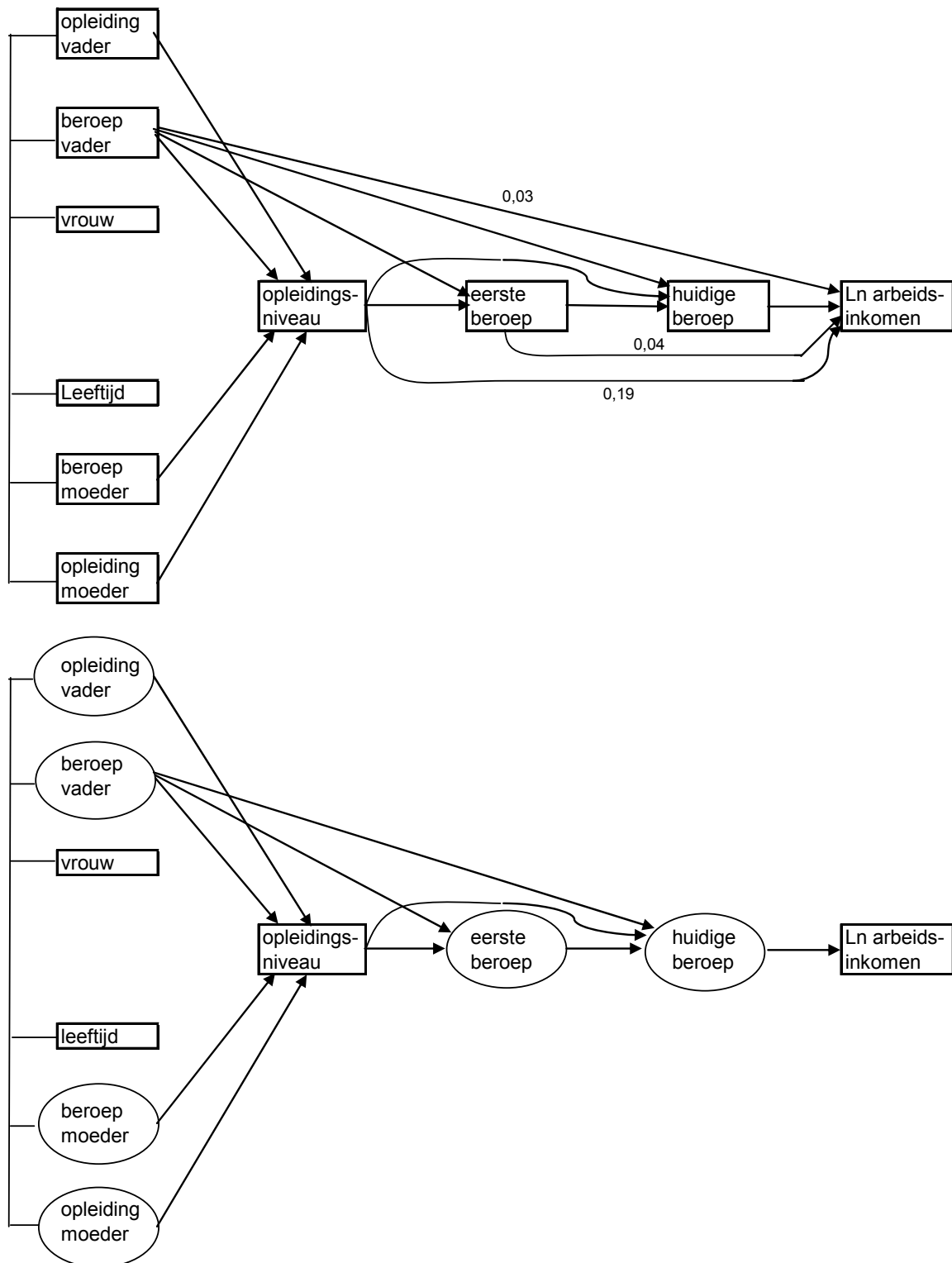
Laat ik een voorbeeld geven over statusverwerving waarbij ik gebruik maak van analyses die zijn uitgevoerd door de geachte collega's De Vries en Ganzeboom (2008). Uit economische theorieën kan men de hypothese afleiden dat het opleidingsniveau een positief direct effect heeft op het loon, omdat hoger opgeleiden productiever zijn in dezelfde functie dan lager opgeleiden en die hogere productiviteit beloond zal worden door de werkgever. Een alternatieve hypothese is dat de lonen grotendeels worden gereguleerd via CAO's waarin lonen direct zijn gekoppeld aan functies.

Daarom zal de invloed van het opleidingsniveau op het loon ook indirect verlopen via de beroepsstatus. Het beroep kan immers worden opgevat als een verzameling soortgelijke functies (Bakker, 1994) met een bijbehorende beroepsstatus. Een andere alternatieve hypothese: als mensen een hoger opleidingsniveau hebben dan voor het beroep nodig is, hebben ze te weinig uitdaging, raken zij teleurgesteld en ongelukkig en worden daarom minder productief (Berg, 1970; Duncan & Hoffman, 1981; Hartog, 2000). Deze hypothesen kunnen worden getoetst aan de hand van de analyses van De Vries en Ganzeboom. In een statusverwervingsmodel laten zij zien dat het corrigeren voor meetfouten leidt tot significant andere conclusies.

In hun onderzoek wordt het beroep van de respondent, diens vader en diens moeder op twee manieren gemeten. De eerste manier is gesloten: mensen delen zichzelf en hun vader en moeder in in een van negen vooraf gegeven beroepscategorieën, zoals ‘geschoolde en leidinggevende handarbeid (bv. automonteur, ploegbaas, elektricien)’. De tweede manier is een open vraag waarbij op grond van de exacte beroepstitel het beroep wordt gecodeerd. De resulterende scores van de open en gesloten vragen zijn omgezet in een score voor de sociaaleconomische status van het beroep.

Eerst is een statusverwervingsmodel geschat waarbij alleen de open vraagstelling is gekozen en er dus niet voor meetfouten wordt gecorrigeerd (bovenste deel figuur 8). De effecten van leeftijd en geslacht en alle niet-significante effecten zijn niet in de figuur ingetekend. Als niet wordt gecorrigeerd voor meetfouten, heeft het opleidingsniveau een significant direct effect van 0,19 op het uurloon. Ook hebben het beroep van de vader en het eerste beroep kleine maar significant direct effecten op het uurloon. Als wel voor meetfouten wordt gecorrigeerd, zijn deze effecten niet langer significant (onderste deel figuur 8). De hypothese dat hoger opgeleiden in dezelfde functie als lager opgeleiden een hoger uurloon hebben omdat ze productiever zouden zijn, wordt nu verworpen. De invloed van opleidingsniveau verloopt geheel indirect via de status van het huidige beroep. Ik hoop u hiermee overtuigt te hebben dat het belangrijk is om naar de onbetrouwbaarheid van variabelen te kijken en daarvoor te corrigeren.

Figuur 8. Een statusverwervingsmodel met en zonder correctie voor meetfouten



De over- en onderdekking bij gebruik van de Gemeentelijke Basisadministratie.

Eerder heb ik verteld dat er onder- en overdekking kan zijn in onderzoek waarbij enquêtes of registers worden gebruikt. Overdekking wil zeggen dat er mensen in het onderzoek opgenomen worden die niet tot de doelpopulatie behoren. Onderdekking wil zeggen dat er mensen worden gemist die wel tot de doelpopulatie behoren. Ik zal nu een voorbeeld uitwerken van onderzoek waarbij de GBA, zeg maar de bevolkingsregisters van de Nederlandse gemeentes, wordt gebruikt om de bevolking van Nederland af te bakenen. Het is goed om te weten dat je verplicht bent om je in te schrijven als je langer dan 4 maanden achter elkaar in Nederland woont of van plan bent dat te doen.

Veel statistieken van het CBS hebben als doelpopulatie de bevolking van Nederland op een bepaald moment of in een bepaalde periode. Vaak wordt daarbinnen nog een subpopulatie onderscheiden, bijvoorbeeld op basis van leeftijd. Zo is de bevolking van Nederland van 15 jaar en ouder een doelpopulatie die veel voorkomt. Ik ga nu de omvang van over- en onderdekking schatten op basis van een groot aantal bronnen. Twee heel belangrijke zijn de GBA en het SSB. Omdat voor 1 januari 2006 veel informatie beschikbaar is, doen we een vingeroefening voor die peildatum.

Op 1 januari 2006 waren er ongeveer 16,334 miljoen mensen in de GBA ingeschreven. Welke mensen verblijven hier nog meer? En welke mensen staan wel in de GBA ingeschreven, maar zijn voor kortere of langere tijd weg? Om te beoordelen welke categorieën wel en niet tot de bevolking van Nederland willen rekenen moeten we ook nog een regel afspreken. Voor het voorbeeld reken ik even alle personen mee die in Nederland wonen en naar verwachting langer dan een maand achtereen in Nederland verblijven.

De eerste categorie is de grensarbeid. *Grensarbeiders* zijn bijvoorbeeld mensen die in België en Duitsland wonen en in Nederland werken. Zij verblijven iedere werkdag voor kortere tijd in Nederland. De inkomende grensarbeid vanuit België bedroeg op 1 januari 2006 36 duizend mensen, en vanuit Duitsland was dat 25 duizend mensen. Dat kunnen zowel werknemers als zelfstandigen zijn. Andersom zijn er ook inwoners van Nederland die in het buitenland werken. De schattingen daarvan zijn wat minder betrouwbaar, maar we komen uit op ongeveer 15 duizend.

Tabel 1

Potentiële bevolking van Nederland op 1 januari 2006

	schatting x1000	plus	min	Onder-/ overdekking?
Ingeschreven in de GBA ¹⁾	16334,2			
<i>Grens- en pendelarbeid</i>				
wonend in België, werknemer in Nederland ²⁾	35,1	35,1		nee
wonend in België, zelfstandige in Nederland ²⁾	1,2	1,2		nee
wonend in Duitsland, werknemer in Nederland ²⁾	24,2	24,2		nee
wonend in Duitsland, zelfstandige in Nederland ²⁾	0,5	0,5		nee
wonend in Nederland, werkend in Duitsland of België ²⁾	14,9		14,9	nee
werknemers uit een ander buitenland dan België en Duitsland ²⁾	64,0	64,0		onderdekking
(tijdelijk) in buitenland werkende inwoners van Nederland ³⁾	29,3		29,3	overdekking
<i>Onderwijs en kinderopvang</i>				
Wonend in België, op school in Nederland ⁴⁾	1,2	1,2		nee
Wonend in Duitsland, op school in Nederland ²⁾	0,7	0,7		nee
Wonend in Nederland, op school in België ⁴⁾	9,8		9,8	nee
Wonend in Nederland, op school in Duitsland ⁵⁾	<0,5		<0,5	nee
Tijdelijke scholieren en studenten in Nederland ²⁾	25,0	25,0		onderdekking
Tijdelijke scholieren en studenten in buitenland	onbekend		onbekend	overdekking
Wonend in België, op kinderdagverblijf in Nederland ⁵⁾	<0,5	<0,5		nee
Wonend in Duitsland, op kinderdagverblijf in Nederland ⁵⁾	<0,5	<0,5		nee
Wonend in Nederland, op kinderdagverblijf in België ⁵⁾	<0,5		<0,5	nee
Wonend in Nederland, op kinderdagverblijf in Duitsland ⁵⁾	<0,5		<0,5	nee
<i>Toeristen en zakenlieden</i>				
Toeristen en zakenlieden op reis in Nederland ⁶⁾	37,8	37,8		nee
Nederlandse toeristen in buitenland ⁷⁾	233,3		233,3	nee
Nederlandse zakenlieden op reis in buitenland	onbekend		onbekend	nee
<i>Niet registreren</i>				
Diplomaten en militairen zonder inschrijvingsplicht ³⁾	5,9	5,9		onderdekking
Asielzoekers ⁸⁾	5,6	5,6		onderdekking
Nederlandse in het buitenland verblijvende gepensioneerden	onbekend		onbekend	overdekking
Buitenlandse in Nederland verblijvende gepensioneerden	onbekend	onbekend		onderdekking
Illegale vreemdelingen ⁹⁾	129,0	129,0		onderdekking
<i>Administratieve vertraging</i>				
Sterfte ¹⁾	0,6		0,6	overdekking
Emigratie ¹⁾	1,2		1,2	overdekking
Geboorte ¹⁾	1,9	1,9		onderdekking
Immigratie ¹⁾	5,1	5,1		onderdekking
Totaal onderdekking		236,5		
Totaal overdekking			31,1	

Bronnen: 1) CBS GBA; 2) CBS SSB; 3) CBS Statistiek Onverzekerden; 4) OCW, 2007; 5) Expert guess; 6) CBS Inkomend Toerisme; 7) CBS Vakantieonderzoek; 8) CBS Asielstatistiek; 9) Van der Heijden et al., 2006

Een tweede groep die we onderscheiden zijn de zogenaamde *pendelmigranten*. Dat zijn mensen die hier voor korte tijd komen werken, maar die geen vast woonadres in Nederland hebben en zich niet hoeven in te schrijven in de GBA. De grens daarvoor ligt bij vier maanden. Op 1 juli 2006 waren er 64 duizend mensen die vanuit een ander buitenland dan België of Duitsland hier een baan hadden als werknemer of zelfstandige. Andersom zijn mensen die in Nederland wonen en tijdelijk in het buitenland werken ook pendelmigranten. Daarvan is de schatting uit het SSB dat het gaat om 30 duizend mensen.

Studenten en scholieren die tijdelijk in Nederland verblijven voor een studie of opleiding, en studenten en scholieren die in België of Duitsland wonen en in Nederland voortgezet of hoger onderwijs volgen, vormen een volgende categorie die niet in de GBA ingeschreven hoeven te staan. Met uitzondering van het primair onderwijs kunnen we daar op grond van de onderwijsstatistieken een goede schatting van maken. Dan gaat het in totaal om ongeveer 27 duizend studenten en scholieren, waarvan het merendeel in het voortgezet en hoger onderwijs. Voor het primair onderwijs weten we het niet precies, maar je mag aannemen dat het dan gaat om leerlingen uit België en Duitsland. Dat aantal is beperkt. Andersom zijn er in Nederland ingeschreven studenten die tijdelijk in het buitenland studeren. Hoeveel dat er precies zijn weten we niet. We weten wel dat er ongeveer 10 duizend scholieren en studenten in Nederland wonen en in België of Duitsland naar school gaan.

We gaan naar nog een jongere categorie: kinderen in kinderdagverblijven. Daar zal ongetwijfeld ook wel wat grensoverschrijding bij te pas komen, maar er zijn geen exacte cijfers over. Mensen met verstand van zaken meldden wel dat de omvang ervan beperkt is. We hebben daarom alle categorieën die je zou kunnen onderscheiden op minder dan een half duizend gezet.

Toeristen en zakenlieden op zakenreis in Nederland verblijven wel voor kortere tijd in Nederland maar zijn niet ingeschreven in de GBA. Op basis van het verblijf in logiesaccommodaties waren er op 1 januari 2006 naar schatting 38 duizend van dergelijke mensen aanwezig (CBS StatLine, 2009). Uit de vakantieonderzoeken blijkt dat er op dat peilmoment ongeveer 233 duizend mensen die in Nederland zijn ingeschreven op vakantie zijn in het buitenland (CBS StatLine, 2009). Hoeveel Nederlandse

zakenlieden op reis zijn in het buitenland op de peildatum is helaas onbekend.

Voorheen werd in de Algemene Volkstellingen wel de categorie ‘varende en rijdende bevolking’ onderscheiden. Sinds de invoering van de GBA op 1 oktober 1994 is ook deze bevolkingscategorie verplicht om zich in te schrijven bij een gemeente. Als ze geen vaste woon- of verblijfplaats hebben dan worden ze ingeschreven in het bevolkingsregister van de gemeente 's-Gravenhage.

Niet-Nederlandse diplomaten en niet-Nederlandse NAVO-militairen die in Nederland wonen, mogen zelf kiezen of ze zich inschrijven in het bevolkingsregister. Een deel blijft daarom ingezetene van het land dat ze vertegenwoordigen. Dat geldt ook voor de meereizende familieleden. Naar schatting gaat het om ongeveer 6 duizend mensen die hier wel wonen en werken, maar niet zijn ingeschreven in de GBA.

Asielzoekers die zes maanden of korter in de centrale opvang verblijven en nog geen verblijfsvergunning hebben, hoeven zich ook niet in te schrijven. In 2005 werden 12350 en in 2006 14470 asielverzoeken ingediend (CBS StatLine, 2009). Op basis van gegevens van 2007 en 2008 schatten we dat ongeveer 83% daarvan een eerst asielverzoek betreft. Als we ervan uitgaan dat asielzoekers die voor het eerste een asielverzoek indienen gelijkmatig over het jaar binnenkomen en zich gemiddeld na een half jaar inschrijven in de GBA, is de onderdekking ongeveer de helft. Dan komen we uit op ongeveer 5,6 duizend op 1 januari 2006.

Een deel van de *gepensioneerden* die in Nederland ingeschreven zijn, verblijven voor een deel van het jaar in het buitenland. Dit worden wel pensionado's genoemd. Andersom kunnen ook in het buitenland geregistreeerde gepensioneerden een deel van het jaar in Nederland verblijven. Er is niet bekend om hoeveel mensen het gaat, en er is zeker ook sprake van overlapping met de toeristen. We nemen daarom noodgedwongen maar aan dat ze bij de onder- en overdekking niet veel gewicht in de schaal zullen leggen.

Illegale vreemdelingen. Ten slotte hebben we een schatting van het aantal vreemdelingen dat gedurende de periode van april 2005 tot 1 april 2006 illegaal in Nederland verbleef. Van der Heijden et al. (2006) schatten dat aantal op 129 duizend. Deze schatting heeft een grote onzekerheidsmarge: als ondergrens geven de auteurs 74 duizend en als bovengrens 184 duizend

illegalen aan. Ongeveer 9 duizend van hen bevinden zich in een nog niet afgeronde asielprocedure. Dat komt nog bovenop de asielzoekers die ik zojuist heb genoemd.

Uiteraard kan er ook sprake zijn van onder- en overdekking door *administratieve vertraging* als we gebruik maken van de GBA om de populatie af te bakenen. Er is sprake van overdekking als mensen nog in de GBA geregistreerd staan terwijl ze feitelijk al zijn geëmigreerd of overleden. Het te laat registreren van geboorte en immigratie leidt tot onderdekking. We hebben de administratieve vertraging in de GBA na zes weken berekend, omdat de gegevens over de omvang van de bevolking van Nederland na zes weken worden vastgesteld. De administratieve vertraging van sterfte na zes weken (0,4%) zorgt ervoor dat 600 sterfgevallen nog niet zijn geregistreerd. Bij emigraties hebben we hetzelfde gedaan: na zes weken is 0,9% nog niet verwerkt. Hier is echter een bijzonder probleem aan de hand. De emigratiedatums zijn dikwijls onbekend en worden daarom op enkele dagen voor de ontdekking van de emigratie gezet. Hier zie je dat ook weer meetfouten optreden: de administratieve vertraging wordt sterk onderschat. Voor wat het waard is: na zes weken zijn 1200 emigraties nog niet in het register opgenomen¹. Ook dit zorgt voor overdekking.

Andersom geldt dat er sprake is van onderdekking door het te laat registreren van geboortes en immigraties. Na zes weken is 99,0% van de geboortes in het register opgenomen. Dat betekent een onderdekking van ongeveer 1900 mensen. Van de immigraties is na zes weken 96,0% administratief verwerkt. Dat zorgt voor een onderdekking van ongeveer vijfduizend mensen. Ook hier is echter een kanttekening op zijn plaats: de immigratiedatum wordt niet altijd adequaat genoteerd. Soms wordt de eerste formele verblijfsdag als datum van immigratie genomen, soms de datum van binnenkomst. Dat zou betekenen dat deze 5 duizend een onderschatting is van de daadwerkelijke onderdekking².

¹ Voor de berekening zijn ook de administratieve afvoeringen als emigratie beschouwd.

² Voor de berekening zijn ook administratieve opvoeringen als immigratie opgevat.

Hoeveel zijn dat er nu *in totaal*? We kunnen alle mensen die ik hierboven heb genoemd optellen om te komen tot een schatting van de omvang van de bevolking in Nederland, en de onder- en overdekking daarvan als je de GBA gebruikt als basis voor de populatie. De onderdekking is dan 237 duizend. Daar komen mogelijk het aantal buitenlandse pensionado's waarvan we niet weten hoeveel er daarvan zijn en de onderschatte administratieve vertraging in de immigratie nog bij. De mensen die voor een klein deel niet aan de definitie voldoen omdat ze te kort in Nederland verblijven, moeten daarvan worden afgetrokken. Hoe groot dat aantal is, is niet bekend. De schatting van de overdekking is 31 duizend. Daar komen de onbekende aantallen tijdelijke scholieren, studenten, pensionado's die tijdelijk in het buitenland verblijven en de onderschatte administratieve vertraging in de emigratie nog bij.

Per saldo zou de bevolking in Nederland ongeveer 200 duizend hoger liggen dan als je de GBA-geregistreerden op 1 januari 2006 telt: geen 16,3, maar 16,5 miljoen. De onzekerheidsmarges zijn echter groot. Alleen al vanwege de marges bij de schatting van de illegalen kan dit 55 duizend hoger of lager liggen. Ook andere schattingen hebben grote marges. Bovendien is de omvang van een aantal categorieën onbekend. Nader onderzoek moet meer inzicht bieden in de marges en de lacunes verder invullen.

Vangst-hervangst-methode

Voor het schatten van de omvang van populaties zoals we dat hierboven hebben gedaan, kun je ook andere methodes gebruiken. Deze behoren tot de familie van de vangst-hervangst-methodes. De vangst-hervangst-methodes zijn ontwikkeld in de biologie om de omvang van populaties dieren of planten te schatten. Voor ons doel zijn de methodes met meer en minder succes in Nederland zijn toegepast. Ik bespreek de meest eenvoudige vorm ervan.

Stel ik wil de populatie teken op een hectare grond in de Amsterdamse waterleidingduinen schatten. Ik baken een stuk grond af ter grootte van een hectare met behulp van schotten. Om de omvang N van de populatie te schatten trek ik eerst de steekproef n_1 van 100 teken. Ik merk de teken die ik getrokken heb en zet deze terug tussen de schotten. Ik wacht even totdat de teken zich goed hebben gemengd met de andere teken. Vervolgens trek

ik een tweede steekproef n_2 van weer 100 teken en ik ga na hoeveel teken in deze steekproef ook al in de eerste steekproef voorkwamen. Deze noemen we m_2 . Dat blijken er 14 te zijn. Als we aannemen dat m_2 een zuiver beeld geeft van de doorsnede van de eerste en de tweede steekproef dan geldt, en dat moet u maar even van me aannemen, dat de populatieomvang geschat kan worden door:

$$\frac{n_1 \times n_2}{m_2} = \frac{100 \times 100}{14} = 714$$

Dat zou betekenen dat er naar schatting ongeveer 714 teken op deze hectare duin aanwezig zijn. In de praktijk is het allemaal veel lastiger dan uit deze eenvoudige formule blijkt. Zo kunt u zich misschien voorstellen dat er teken worden geboren en sterven, of dat er teken over de schotten springen, of dat teken leren van de eerste keer dat ze gevangen zijn en zich niet voor een tweede keer laten vangen of zich juist heel graag voor een tweede keer laten vangen, of dat niet gezien wordt dat een teek gemerkt is in de eerste vangst. Dit heeft een verstorende invloed op de schatting van het populatietotaal. Samen met Van der Heijden en Cruijff van de Universiteit Utrecht en CBS-collega's Oudhof en Van der Vliet gaan we deze methode verder ontwikkelen en toepassen voor het regulier schatten van de totale Nederlandse bevolking. Belangrijk blijft echter ook om aan te geven of de onder- en overdekking selectief is, en als dat zo is, naar welke kenmerken dat het geval is. We kunnen dan veel beter aangeven wat de vertekening is die door onder- en overdekking wordt veroorzaakt. Daarvoor is het onderzoek naar welke categorieën nu de over- en onderdekte populatie vormen nuttig.

Verbeteren van de samenwerking tussen universiteit en CBS

Naast zelf onderzoek doen wil ik nog een paar dingen bereiken. Ik wil het onderwerp van de leerstoel op de agenda zetten van de Nederlandse methodologen en de samenwerking tussen universiteiten en het CBS verbeteren. De schatting van de omvang van de Nederlandse bevolking is daarvan een voorbeeld, maar ook op andere terreinen wil ik dat bevorderen.

De verwerking van registerinformatie tot bruikbaar materiaal voor statistisch onderzoek is tot op heden vooral een taak geweest van de

statistische bureaus, en dan in het bijzonder die uit de Scandinavische landen en Nederland. De publicaties van Statistics Denmark (1995), Al en Bakker (2000), Schulte Nordholt, Hartgers en Gircour (2004), Statistics Finland (2004), CBS (2006), en Wallgren en Wallgren (2007), zijn recente beschrijving van de werkwijze, 'best practices' en producten. Inmiddels sluiten steeds meer statistische bureaus in de rij aan³. De universitaire onderzoekers hebben er slechts mondjesmaat aan meegedaan. Toch hebben ook universitaire onderzoekers er alle belang bij om de kwaliteit van registerinformatie te bepalen, al was het maar omdat deze steeds vaker wordt gebruikt in wetenschappelijk onderzoek.

Hoewel onderzoekers van de statistische bureaus en uit de universitaire wereld een soortgelijke basistraining hebben genoten, verschilt hun professionele ontwikkeling sterk. Ontwikkelen statistici van de statistische bureaus vooral methoden op het terrein van waarneming en verwerking van gegevens, de statistici van de universiteiten concentreren zich veel vaker op de ontwikkeling en toepassing van geavanceerde analysetechnieken. Dat heeft twee gevolgen. Het eerste gevolg is dat de statistici van de statistisch bureaus afstand voelen tot de universitaire onderzoekers, waardoor het lang duurt voordat zij de nieuw ontwikkelde methoden uit de universitaire wereld toepassen, als dat al gebeurt. En het tweede gevolg is dat de universitaire onderzoekers vaak geen idee hebben welke behoefte bestaat aan methodeontwikkeling ten behoeve van de officiële statistiek (Grünwald en Körner, 2005).

Het bij elkaar brengen van de verschillende disciplines heeft als meerwaarde dat gezamenlijk de kwaliteit van de methoden van waarneming via registers en de analyses daarop verhoogd kan worden. Dit is profijtelijk voor de officiële statistiek en voor het universitair onderzoek dat gebaseerd wordt op registerinformatie. Naar verwachting is zal dat laatste een steeds grotere vlucht zal nemen door de toegenomen mogelijkheden die onder meer door het CBS daarvoor worden geboden (Dykstra, 2006).

³ Een interessante nieuwkomer is België. In 2001 heeft dat land nog een traditionele volkstelling gehouden, maar voor 2011 worden de gegevens mogelijk ontleend aan bestaande registraties. Daarbij wordt onder meer gebruik gemaakt van de gegevens van de zogenaamde kruispuntbank. Zie bijvoorbeeld Tielens en Stimpson (2004) en Vermandere (2004).

Concluderend zeg ik: trek alle registers open, maar doe dat met beleid en met oog voor de methodologische valkuilen en de soms gebrekkige kwaliteit van registerdata.

Nawoord

Dames en heren, aan het einde van mijn rede gekomen, wil ik de directie van het Centraal Bureau voor de Statistiek en het College van Bestuur van de Vrije Universiteit bedanken voor het instellen van de leerstoel en mijn benoeming als hoogleraar daarop. In het bijzonder dank ik Wim van Nunspeet, Harry Ganzeboom en Bert Klandermans hiervoor. Frans Leeuw dank ik voor het opperen van het instellen van een leerstoel registermethodologie in de CBS Adviesraad Economie en Maatschappij.

Het is een grote uitdaging om dit nog zo prille vakgebied verder te ontginnen. Dat kan ik niet alleen. Ik hoop dat ik een brug kan slaan tussen de methodologen die werkzaam zijn bij mijn vakgroep, maar ook anderen binnen de universitaire wereld, en de methodologen en onderzoekers van het CBS. We kunnen niet zonder elkaar en door samen te werken kunnen we de puzzelstukjes op hun plek laten vallen.

De twee werelden van het CBS en de universiteit zijn gedurende mijn carrière afwisselend van groot belang geweest voor mijn ontwikkeling. Ik beschouw Jaap Dronkers als mijn belangrijkste leermeester op wetenschappelijk terrein. In de vele projecten die wij aan het begin van mijn carrière samen hebben uitgevoerd, bracht hij mij een wetenschappelijke instelling bij die mij nog steeds van pas komt. Ik heb ook veel geleerd van de discussies in de overleggroepen Longitudinaal School- en Beroepsloopbaanonderzoek en Sociale Stratificatie, tegenwoordig verenigd in Sociale Ongelijkheid en Levenslopen. Ook binnen het CBS zijn er velen die aan mijn ontwikkeling hebben bijgedragen. Het CBS is een mooi bedrijf met getalenteerde medewerkers, waar ruimte is voor het nadenken over methodologische vernieuwing. Ik ben er trots op dat ik aan die methodologische vernieuwing mede vorm mag geven. Het is wel grotendeels gericht op het produceren van nieuwe reguliere statistieken en wijkt daarin af van de universitaire wereld.

Mam, je hebt inmiddels de gezegende leeftijd van eenennegentig bereikt. Ik vind het geweldig dat ik dit nog met je kan delen. Dank je voor de bagage die je mij meegaf.

Trek alle registers open!

Christianne, Floor en Jette, Gabriëlle en Tobias, wij zijn samen het bewijs van de feilbaarheid van administratieve gegevens: de afstand tussen jullie en mij is veel kleiner dan je uit de bevolkingsregisters zou denken. Dank jullie voor alles.

Ik heb gezegd.

Referenties

- Al, P. & B.F.M. Bakker, 2000, Re-engineering Social Statistics by micro-integration of different sources. An introduction. In: P. Al & B.F.M. Bakker (red.), Re-engineering social statistics by micro-integration of different sources. Themanummer *Netherlands Official Statistics*, jrg. 15, summer, blz. 4-6
- Arts, K., B.F.M. Bakker & E. van Lith, 2000, Linking administrative registers and household surveys. In: P. Al & B.F.M. Bakker (red.), Re-engineering social statistics by micro-integration of different sources. Themanummer *Netherlands Official Statistics*, jrg. 15, summer, blz. 16-22
- Bakker, B.F.M., 1994, De CBS Standaard Beroepenclassificatie 1992, In: *Tijdschrift voor Arbeidsvraagstukken*, jrg. 10, nr. 4, blz. 322-335
- Bakker, Bart F.M., 2009, *Micro-integratie. CBS Methodenreeks*. (Den Haag/ Heerlen: Centraal Bureau voor de Statistiek)
- Bakker, B.F.M., & J. van Rooijen, 2000, One figure for the supply and demand of services. In: P. Al & B.F.M. Bakker (red.), Re-engineering Social Statistics by micro-integration of different sources. Themanummer *Netherlands Official Statistics*, jrg. 15, summer, blz. 40-46
- Belson, W.A., 1986, *Validity in survey research* (Brookfield: Gower)
- Berg, I., 1970, *Education and jobs. The great training robbery* (New York: Praeger)
- Besjes, G. & R. van Gaalen, 2008, Jong geleerd, fout gedaan? Een onderzoek naar de intergenerationele overdracht van criminaliteit met gegevens uit het Sociaal Statistisch Bestand. In: *Bevolkingstrends*, jrg. 56, nr. 2, blz. 23-31
- Blom, M., J. Oudhof, R.V. Bijl & B.F.M. Bakker (red.), 2005, *Verdacht van criminaliteit. Allochtonen en autochtonen nader bekeken. WODC-cahier 2005-2* (Den Haag/Voorburg: WODC/CBS)
- Borghans, L., 2007, Anders denken. In: *E-data&research*, jrg. 2, nr 2, blz. 8
- Bound, J., C. Brown, G.J. Duncun & W.L. Rodgers, 1994, Evidence on the validity of cross-sectional and longitudinal labor market data. In: *Journal of Labor Economics*, jrg. 12, nr. 3, blz. 345-368

- Brick, J.M., R. McGuinness, S.J. Lapham, M. Cahalan, D. Owens & L. Gray, 1995, Interviewer variance in two telephone surveys. In: American Statistical Association: *Proceedings of the Section on Survey Research Methods*, blz. 447–452
- CBS, 1995, *Van cijferfabriek naar knooppunt op de elektronische snelweg* (Voorburg/Heerlen: Centraal Bureau voor de Statistiek)
- CBS, 2006, *Sociale samenhang in beeld. Het SSB nu en straks* (Voorburg/Heerlen: Centraal Bureau voor de Statistiek)
- CBS StatLine, 2009, *StatLine* (Den Haag/ Heerlen: Centraal Bureau voor de Statistiek)
- Duncan, G.J. & D.H. Hill, 1985, An investigation of the extent and consequences of measurement error in labor economic survey data. In: *Journal of Labor Economics*, jrg. 3, nr. 4, blz. 508-532
- Duncan, G. & S.D. Hoffman, 1981, The incidence and wage effects of overeducation. In: *Economics of Education Review*, jrg. 1, nr. 1, pp. 75-86
- Dykstra, P., 2006, Scandinavische toestanden in Nederland: toegang tot register-bestanden. In: *Mens en Maatschappij*, jrg. 81, nr. 3, blz. 198-201
- I. Fellegi & A. Sunter, 1969, A theory of record linkage. In: *Journal of the American Statistical Association*, jrg. 64, blz. 1183-1210
- Ganninger, M., S. Häder & S. Gabler, 2007, *Design effects and interviewer effects in the European Social Survey: Where are we now and where do we want to go tomorrow?* (Mannheim: Centre for Survey Research and Methodology)
- Groves, R.M., F.J. Fowler jr., M.P. Couper, J.M. Lepkowski, E. Singer, & R. Tourangeau, 2004, *Survey Methodology* (New York: Wiley Interscience)
- Grünewald, W. & T. Körner, 2005, Quality on its way to maturity: results of the European conference on Quality and methodology in official statistics (Q2004). In: *Journal of Official Statistics*, jrg. 21, nr. 4, blz. 747-759
- Hartog, J., 2000, Over-education and earnings: where are we, where should we go? In: *Economics of Education Review*, vol. 19, pp. 131-147

Heijden, P.G.M., G. van Gils, M. Cruijff en Dave Hessen, 2006, *Een schatting van het aantal in Nederland verblijvende illegale vreemdelingen in 2005* (Utrecht: IOPS Universiteit Utrecht)

Hooghiemstra, E., 2003, *Trouwen over de grens: achtergronden van partnerkeuze van Turken en Marokkanen in Nederland* (Den Haag: Sociaal en Cultureel Planbureau)

Jöreskog, K. & D. Sörbom, 1996, *LISREL 8. User's reference guide* (Chicago: Scientific Software International)

Kapteyn, A. & J.Y. Ypma, 2007, Measurement error and misclassification: a comparison of survey and administrative data. In: *Journal of Labor Economics*, jrg. 25, nr. 3, blz. 513-551

Kim, B. & G. Solon, 2005, Implications of mean-reverting measurement error for longitudinal studies of wages and employment. In: *Review of Economics and Statistics*, jrg. 87, nr. 1, blz. 193-196

Kline, R.B., 2005, *Psychological testing: a practical approach to design and evaluation* (New York: SAGE)

Kroese, B., R.H. Renssen en M. Trijssenaar, 2000, Weighting or imputation: constructing a consistent set of estimates based on data from different sources. In: In: Pieter Al en Bart F.M. Bakker (eds.), *Re-engineering Social Statistics by micro-integration of different sources*. Themanummer *Netherlands Official Statistics*, jrg. 15, nr. summer blz. 23-31

Leeuw, E. de, & W. de Heer, 2002, Trends in household survey nonresponse: a longitudinal and international comparison. In: R.M. Groves, D. A. Dillman, J.L. Eltinge & R.J.A. Little (red.), *Survey nonresponse* (New York: Wiley), blz. 41-54

Maarseveen, J.G.S.J. van, 1999, Volkstellingen 1795-1971. In: J.G.S.J. van Maarseveen (red.), *Algemene tellingen in de twintigste eeuw* (Voorburg/Heerlen: Centraal Bureau voor de Statistiek)

Mulry, M.H., S.L. Bean, D.M. Baunder, D. Wagner, T. Mule & R. Petroni, 2006, Evaluation of estimates of census duplication using administrative records information. In: *Journal of Official Statistics*, jrg. 22, nr. 4, blz. 655-679

Nobel, J.R., 2003, De deur van het demografisch laboratorium. Perspectieven voor externe benutting van de virtuele volkstelling 2001. In: B. Bakker & L. Putman (red.), *De virtuele Volkstelling en het Sociaal Statistisch Bestand* (Amsterdam: SISWO), blz. 101-121

OCW, Ministerie van Onderwijs, Cultuur en Wetenschappen, 2007, De wederzijdse onderwijsparticipatie van leerlingen en studenten in Nederland en Vlaanderen. Schooljaar 2005/2006 (Brussel/ Den Haag: Ministerie van Onderwijs en Vorming/ Ministerie van Onderwijs, Cultuur en Wetenschappen)

Pannekoek, J., 1988, Interviewer variance in a telephone survey, *Journal of Official Statistics*, jrg. 4, nr. 4, blz. 375-384

Pedace, R., & N. Bates, 2001, Using administrative records to assess earnings reporting in the survey of income and program participation, *Journal of Economic and Social Measurement*, jrg. 26, nrs. 3-4, blz. 173-192

Pischke, J.-S., 1995, Measurement error and earnings dynamics: some estimations from the PSID validation study. In: *Journal of Business and Economic Statistics*, jrg. 13, nr. 3, blz. 305-314

Rodgers, W.L., C. Brown & G.J. Duncan, 1993, Errors in survey reports of earnings, hours worked and hourly wages, *Journal of the American Statistical Association*, jrg. 88, nr. 424, blz. 1208-1218

Scheele, D., R. van Gaalen & J. van Rooijen, 2009, *Werk en inkomen na massaontslag: De zekerheid is niet van de baan. WRR-verkenning V20* (Den Haag: WRR/CBS)

Schulte Nordholt, E., M. Hartgers & R. Gircour (red.), 2004, *The Dutch virtual census of 2001: analysis and methodology* (Voorburg/ Heerlen: Statistics Netherlands)

Sikkel, D., 1988, *Quality aspects of statistical data collection* (Amsterdam: Sociometric Research Foundation)

Statistics Denmark, 1995, *Statistics on Persons in Denmark – A register-based statistical system* (Luxembourg: Eurostat)

Statistics Finland, 2004, *Register based statistics. Best practices* (Helsinki: SF)

Stinson, M., 2002, *Estimating measurement error in SIPP annual job earnings: a comparison of census survey and SSA administrative data* (Suitland: US Census Bureau)

Stoop, I.A.L., 2005, *The hunt for the last respondent. Nonresponse in sample surveys* (Den Haag: SCP)

Tielens, M. & A. Stimpson, 2004, Administratieve bestanden of enquêtes? In: *Over.Werk*, jrg. 14, nr. 3, blz. 88-97

Vermandere, C., 2004, 4 miljoen werkzame Belgen samen op een schijfje. In: *Over.Werk*, jrg. 14, nr. 1-2, blz. 92-97

Verschuren, P.J.M., 1991, *Structurele modellen tussen theorie en praktijk* (Utrecht: Spectrum)

Vries, J. de, & H.B.G. Ganzeboom, 2008, Hoe meet ik beroep? Open en gesloten vragen naar beroep toegepast in statusverwervingsonderzoek. In: *Mens en Maatschappij*, jrg. 83, nr. 1, pp. 70-95

Wallgren, A. & B. Wallgren, 2007, *Register-based statistics: administrative data for statistical purposes. Wiley series in survey methodology* (New York: Wiley)

Zorlu, A. en C.H. Mulder, 2008, Initial and subsequent location choices of immigrants to the Netherlands. In: *Regional Studies*, vol. 42, nr. 2, pp 245-264

Trek alle registers open!

©BAKKER, Faculteit der Sociale Wetenschappen, Vrije Universiteit
Amsterdam, 2009
ISBN: [978-90-9024827-1](#)